Topic Modeling for Untargeted Substructure Exploration in Metabolomics

Justin J.J. van der Hooft[a,d], Joe Wandy[a,b], Michael P. Barrett[a,c], Karl E.V. Burgess[a] and Simon Rogers[a,b,1]

[a]Glasgow Polyomics, University of Glasgow, Glasgow G61 1QH, United Kingdom
[b]School of Computing Science, University of Glasgow, Glasgow G12 8RZ, United Kingdom
[c]Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow G12 8TA, United Kingdom
[d]Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, United Kingdom

November 4, 2016

Abstract

The potential of untargeted metabolomics to answer important questions across the life sciences is hindered due to a paucity of computational tools that enable extraction of key biochemically relevant information. Available tools focus on using mass spectrometry fragmentation spectra to identify molecules whose behavior suggests they are relevant to the system under study. Unfortunately, fragmentation spectra cannot identify molecules in isolation, but require authentic standards or databases of known fragmented molecules. Fragmentation spectra are, however, replete with information pertaining to the biochemical processes present; much of which is currently neglected. Here we present an analytical workflow that exploits all fragmentation data from a given experiment to extract biochemically-relevant features in an unsupervised manner. We demonstrate that an algorithm originally utilized for text-mining, Latent Dirichlet Allocation, can be adapted to handle metabolomics datasets. Our approach extracts biochemically-relevant molecular substructures ('Mass2Motifs') from spectra as sets of co-occurring molecular fragments and neutral losses. The analysis allows us to isolate molecular substructures, whose presence allows molecules to be grouped based on shared substructures regardless of classical spectral similarity. These substructures in turn support putative de novo structural annotation of molecules. Combining this spectral connectivity to orthogonal correlations (e.g. common abundance changes under system perturbation) significantly enhances our ability to provide mechanistic explanations for biological behavior.

Significance Statement

Tandem MS is a common technique for compound identification in untargeted metabolomics experiments. Due to a lack of reference spectra, the majority of molecules cannot be

identified and many spectra cannot be used. We present MS2LDA, an unsupervised method (inspired by text mining algorithms) that extracts common patterns of mass fragments and neutral losses — Mass2Motifs — from collections of fragmentation spectra. Structurally characterized Mass2Motifs can be used to annotate molecules for which no reference spectra exist and expose biochemical relationships between molecules. For 4 beer extracts, without training data, we show that with 30 structurally characterized Mass2Motifs we can annotate approximately 3 times as many molecules as with library matching. These Mass2Motifs were validated in reference spectra from Global Natural Products Social Molecular Networking (GNPS) and Massbank.

Keywords metabolomics; mass spectrometry; fragmentation; bioinformatics; topic modelling

1 Introduction

Mass Spectrometry (MS) based metabolomics aims to capture the entire small molecule composition of biological systems. Analysis of MS metabolomics data is challenging as many molecules cannot be identified from their mass (e.g. isobaric molecules, and isomers) [9, 15, 31]. Separation by liquid chromatography prior to MS (LC-MS) can add discriminatory information but does not solve the problem as isomers can exhibit similar chromatographic behavior, and chromatographic retention time is currently unpredictable.

Fragmentation spectra have been used to partially overcome this problem [8, 14, 18]. Most tools compare individual fragmentation spectra to reference spectra [14, 20] stored in public databases, e.g. MassBank [13] or Human Metabolome Database [33], and are thus constrained by the limited number of reference spectra [1, 7, 21]. Poor identification coverage can result in poor biochemical insight. We propose a method that analyses all acquired fragmentation spectra to expose underlying biochemistry without relying on metabolite identification, inspired by machine learning techniques developed initially for text processing [2].

The paucity of techniques that share information across fragmentation spectra can be explained by the complexity of fragmentation data [10]. One example, 'Molecular Networking', clusters MS1 peaks by their MS2 spectral similarity such that one structurally annotated metabolite in a cluster facilitates structural annotation of its neighbors [32, 34]. However, spectral features causing the clustering must be extracted manually and only MS2 spectra with high overall spectral similarity are grouped. Another package, MS2Analyzer [17] mines MS2 spectra for specific features defined by the user (i.e., mass fragments and neutral losses). Some will be common to many experiments (e.g. CO or H_2O losses), but sample-specific features are easily overlooked. Whilst Molecular Networking requires no user intervention it may fail to group molecules that share small substructures, whilst MS2Analyzer can find all molecules that share a particular set of features provided they are user-specified. Our approach, MS2LDA, which is based on Latent Dirichlet Allocation (LDA) [2], retains the benefits of both of these approaches whilst losing the shortfalls – it can find relevant substructures based on the co-occurrence of mass fragments and neutral losses, and group the molecules accordingly. Although adapted to other domains (e.g. genomics [5] and transcriptomics [22]) LDA has never been used to exploit the parallels between MS2 data and text.

Fragmentation spectra contain recurring patterns of fragments and losses due to common biological substructures (e.g. a hexose unit, or a carboxyl group loss). We assume each observed spectrum is comprised of one or more such substructures, an assumption successfully used in

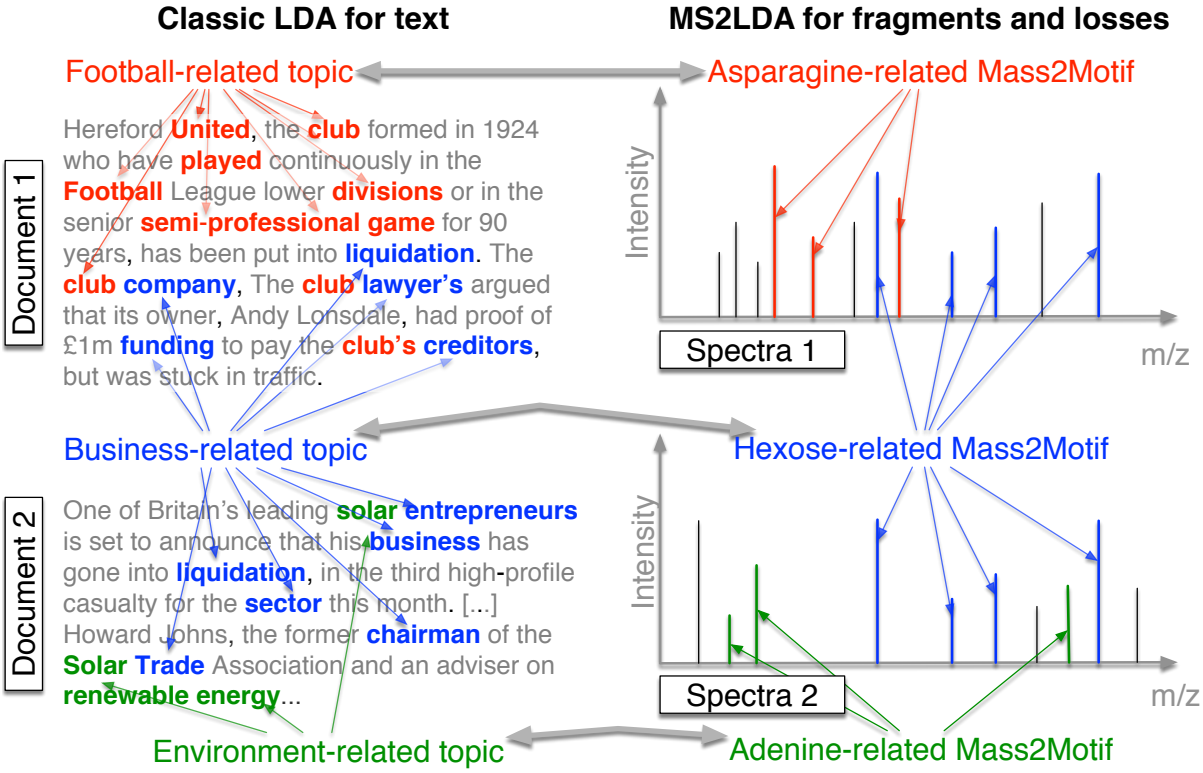


Figure 1: Analogy between LDA for text and MS2LDA. LDA finds topics interpreted as 'football related', 'business-related' and 'environment related'. MS2LDA finds sets of concurring mass fragments or losses (Mass2Motifs) that can be interpreted as 'Asparagine-related', 'Hexose-related' and 'Adenine-related'.

other workflows [8, 27]; however, no unsupervised strategy exists that finds mass fragmental-based substructures without training data. Figure 1 demonstrates the parallels between text and fragmentation data. LDA decomposes documents into topics based on co-occurring words, while MS2LDA decomposes fragmentation spectra into blocks of co-occurring fragments and losses, referred to as 'Mass2Motifs'. Using all of the fragmentation spectra generated by data-dependent mass fragmentation analysis (DDA), MS2LDA learns the conserved substructures (the Mass2Motifs) and the decomposition of the fragmentation spectra into Mass2Motifs.

Our analysis pipeline (see also SI Section 1) performs data pre-processing, extracts Mass2Motifs and allows interactive exploration of the results. Through the analyses of four beer extracts, we show that without labelled training data or metabolite identification, MS2LDA extracts mass patterns indicative of biological substructures that can be structurally annotated, some of which are pathway related. These can aid in the putative de novo annotation or functional classification of otherwise unidentifiable molecules. Many more molecules can be annotated in this way than through comparison with reference spectra. Grouping of molecules based on common substructures is particularly useful for hypothesis-generating research. For example, hypotheses as to the source of variation in metabolite abundances can be obtained by linking MS1 abundance changes to the presence of common substructures.

MS2LDA

Data, in the form of .mzXML (full scan) and .mzML (fragmentation) files, is pre-processed using XCMS [25] and MzMatch [23] for peak detection and RMassBank [26] for detecting MS1-MS2 pairs, before matrix formation by aligning MS2 features across different spectra. The resulting matrix has MS2 features (fragments and losses) as rows, and MS2 peaks as columns. The values in the matrix are the MS2 feature intensities which are subsequently transformed into integer 'counts' (SI Appendix Section S1).

For LDA inference, we have implemented both collapsed Gibbs sampling [11] and Variational Bayes [2] in Python. The output is a set of Mass2Motifs and assignments of Mass2Motifs to each MS1 peak. In addition, we provide an optional elemental formula assignment step [3, 4, 16] to assign candidate elemental formulae to the MS2 features and MS1 peaks. On a laptop (Intel Core i7, 16GB RAM) running the workflow for one beer sample takes around 20 minutes for the feature extractions, and between 30 minutes (Variational Bayes) to 1 hour (Gibbs Sampling) for the inference. The LDA output can be explored in the MS2LDAvis module (customized from LDAvis [24]). Full details are provided in SI Appendix Section S1. We used MS2LDAvis to inspect Mass2Motifs with degree $\geq$10 (i.e. that were present in ten or more spectra) and structurally characterized them (assigned a substructural annotation) at varying levels of confidence (see SI Appendix section S2.1) through expert knowledge and matching of the Mass2Motif spectra to reference spectra in MzCloud (www.mzcloud.org).

2 Results

The MS2LDA workflow was independently applied to 4 beer extracts. After pre-processing, each sample consisted of around 1,000 MS peaks in both positive and negative ionization mode (see SI Appendix Section S2.2). 300 Mass2Motifs were extracted for each data file and checked for biochemical relevance. 30-40 Mass2Motifs in each of the positive ionization mode files were

structurally characterized (see SI Appendix Table S-4) and diverse biochemically relevant substructures found included histidine, phenylalanine, adenine, hexose-units, and structural features such as water or carboxyl group loss.

The degree of Mass2Motifs (the number of spectra in which they occurred) varied from 1 to over 200, demonstrating that MS2LDA can extract both generic and specific structural features. The number of Mass2Motifs within each spectrum also varied (around 600 spectra in each file consisted of one Mass2Motif, 300 of two, 50 of three, and 20 of four or more). Across the four files, an average of 70% of spectra (see SI Appendix Section S2.3) include at least one characterized Mass2Motif, demonstrating the power of MS2LDA for data reduction – i.e. structurally characterizing just 30-40 of the discovered Mass2Motifs provides biochemical insight into 70% of the spectra. For comparison, we matched spectra to the MassBank and National Institute of Standards and Technology libraries (see SI Appendix Section S2.4) at a threshold of 90% normalized score, obtaining hits for only 25% and 6% of the spectra, respectively, demonstrating the wide coverage possible with MS2LDA.

2.1 Automatic, Unsupervised, Chemical Substructure Discovery

Mass2Motifs cover a diverse set of biochemical features, including amino acid related (i.e. histidine, leucine, tryptophan, and tyrosine), nucleotide related (i.e. adenine, cytosine, and xanthine), and other molecules such as cinnamic acid, ferulic acid, ribose and N-acetylputrescine. Mass2Motifs related to the same substructure or structural feature were consistently found across multiple beers (e.g. hexose-related Mass2Motifs were present in all positive ionization mode files). Differences in degree and absence of some Mass2Motifs across the extracts shows that MS2LDA captures variability in metabolic composition.

An example of ferulic acid (a compound present in cereals, an ingredient of beer) is given in Figure 2. Two of the eleven spectra that include Mass2Motif 19 are shown. Conserved mass fragments are clearly visible across the two spectra. Unlike existing software, e.g. MS2Analyzer [17], our method is unsupervised and has no need for prior knowledge about fragments of interest. It is of note that the neutral loss of the complete ferulic acid moiety was also included by MS2LDA, demonstrating that both fragments and losses can be present in a motif. MS2LDA is able to extract a relatively rare biochemically relevant pattern (present in 11 of the spectra), despite the individual spectra being quite different.

Positive ionization mode fragmentation spectra generally provide larger sets of conserved fragments but some Mass2Motifs e.g. those related to phosphate and sulfate groups (fragments at 78.9593 ($[PO_3]^-$) and 79.9575 ($[SO_3]^-$) m/z, respectively) were more easily identifiable in negative mode; an argument to use both ionization modes. Three of the characterized positive mode Mass2Motifs pointed to the highly similar aromatic substructures of phenylethene, cinnamic acid (cinnamate), and phenylethyleneamine (i.e., [phenylalanine – CHOOH]), demonstrating discrimination of very similar yet functionally different substructures (see SI Appendix Section S2.6).

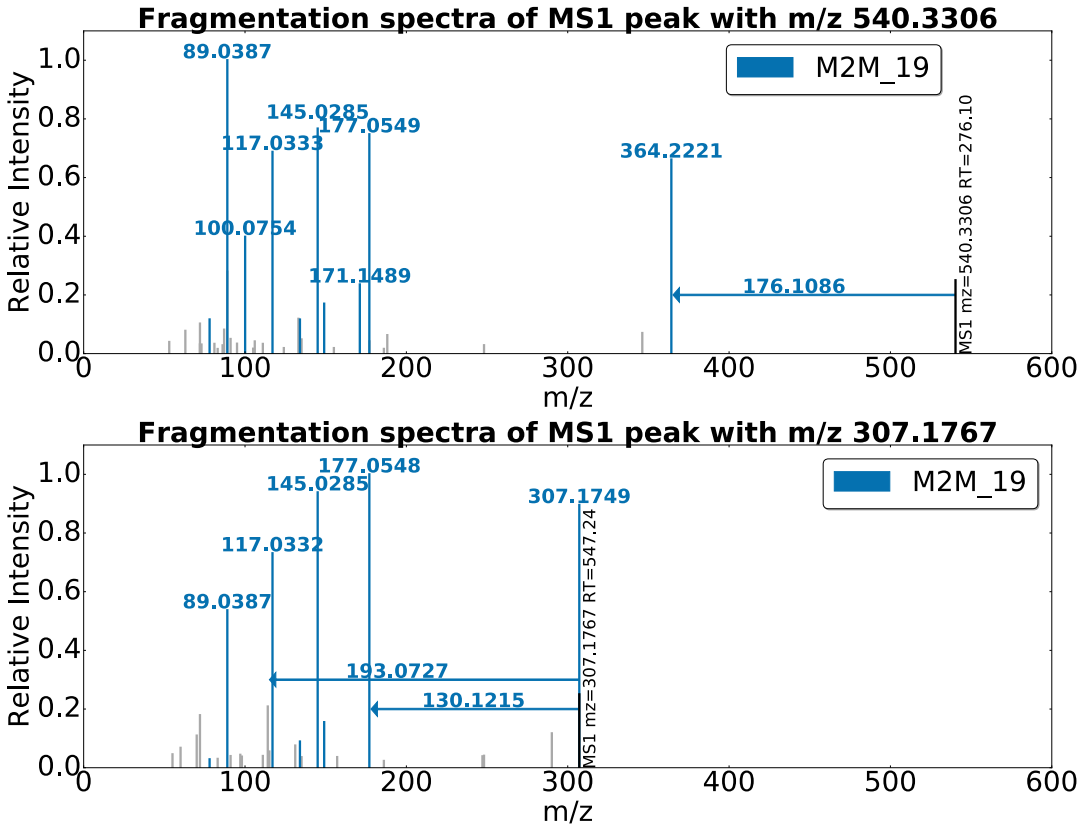


Figure 2: Two spectra, from the beer3 positive ionization mode file, each of which includes Mass2Motif 19, annotated as the plant derived ferulic acid substructure. The mass fragments and neutral losses (arrows originating at the precursor ions) included in Mass2Motif 19 are highlighted in colour. Fragments not explained by Mass2Motif 19 are light grey. The probabilistic nature of MS2LDA means that Mass2Motifs will not necessarily be identical in all spectra in which they appear.

2.2 Structurally Characterized Mass2Motifs Validated in Authentic Standards

Reference molecules in the beer extracts were identified based on chromatographic co-elution and corresponding exact mass. As their identity is known, we can validate our structurally characterized Mass2Motifs. Of the 45 reference molecules we could identify, 38 included one or more characterized Mass2Motifs, 32 of which were validated (i.e. do indeed include the relevant substructure), despite the fact that the Mass2Motif was characterized without a reference molecule identification.

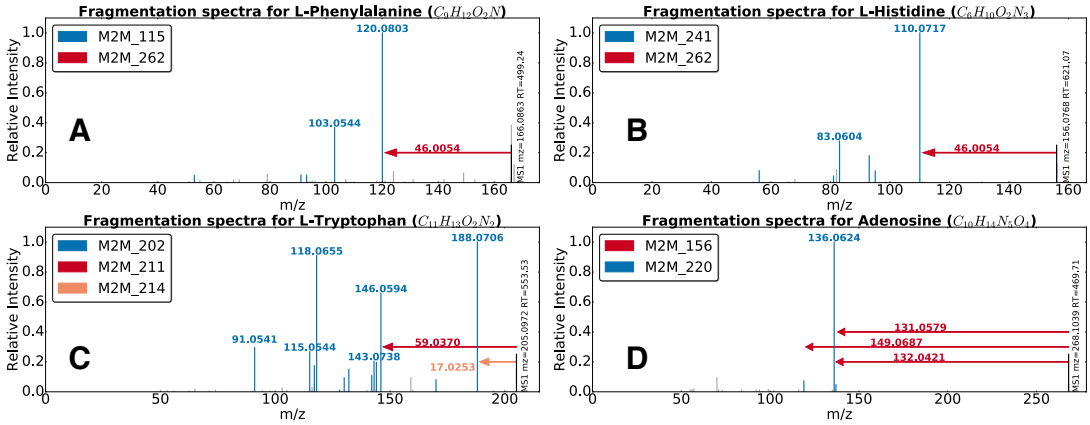


Figure 3: Mass2Motif spectra of identified metabolites **A)** L-histidine, **B)** L-phenylalanine, **C)** L-tryptophan, and **D)** adenosine. Characterized motifs are indicated by color. Full details of the mentioned Mass2Motifs can be found in SI Appendix, Section S2.7.

Some examples are provided in Figure 3. The spectra for phenylalanine (Figure 3A) and histidine (Figure 3B) share Mass2Motif 262, indicating the presence of a free (underivatized) carboxylic acid group. The loss of CHOOH (Mass2Motif 262) is in fact a common characteristic for many other underivatized amino acids and free organic acids and was associated with 10 of the 18 identified amino acids structures (the remaining 8 prefer alternative fragmentation routes – e.g. see the amine loss (Mass2Motif 214) in tryptophan, Figure 3C). The other Mass2Motifs (115, 241) in Figure 3A and B are related to phenylalanine and histidine, respectively (more details in SI Appendix Section S2.7). Figure 3D is the MS2 spectrum of adenosine, which consists of an adenine molecule conjugated to a ribose sugar molecule. The two associated Mass2Motifs (156, 220) represent these two biochemically relevant structural features (i.e., adenine substructure and a ribose sugar loss).

Spectra can include multiple Mass2Motifs. In each of Figures 3A to 3D, we observe two or more Mass2Motifs. We know of no other method that can do this without training spectra consisting of known structures, or prior knowledge of interesting feature combinations. Multiple Mass2Motifs can also explain the same feature in one spectrum, i.e. the fragments 110.0717 ($C_5H_8N_3$, [M+H]+) and 120.0803 ($C_8H_{10}N$, [M+H]+) in Figures 3A and 3B are explained by Mass2Motifs 241 and 115 and also by the 46.0054 loss (CHOOH) of Mass2Motif 262. This demonstrates the manner in which MS2LDA decomposes molecules into their constituent building blocks, allowing for de novo metabolite annotation.

2.3 Mass2Motifs Aid *de-novo* Metabolite Annotation

On average 70% of the fragmented MS1 features are explained by at least one structurally characterized Mass2Motif and can therefore be automatically classified. For comparison, we performed spectral matching using the National Institute of Standards and Technology MS/MS database for small molecules (http://chemdata.nist.gov /mass-spc/msms-search/) and MassBank [13] on 7 of the metabolites annotated via the ferulic acid Mass2Motif. Only 1 returned a ferulic acid related hit, in spite of the clear presence of ferulic acid in all spectra (see e.g. Figure 2). The Mass2Motif itself can be represented as a spectrum and be subjected to spectral matching, resulting in trans-ferulic acid as the best hit (hinting at the possibility of automatic Mass2Motif annotation). Spectra that are explained by the Mass2Motifs related to histidine, tyrosine, and tryptophan were also subjected to spectral matching. From 39 metabolites annotated with help of MS2LDA, 7 resulted in correct hits with another 8 producing structurally related hits (see SI Appendix Section S2.4). These results clearly demonstrate the annotative power of MS2LDA, through which annotations can be made by matching only small portions of the spectra and therefore allowing annotation (classification) of molecules not present in databases. In summary, our experiments show that MS2LDA is able to annotate approximately three times as many metabolites as spectral matching. In addition, MS2LDA can annotate and group spectra based on neutral losses (e.g. the loss of CHOOH), which is not possible with spectral matching.

To further assess the use of the structurally characterized Mass2Motifs in metabolite annotation, we used MS2LDA to decompose 1953 and 5670 spectra from MassBank and the Global Natural Products Social Molecular Networking (GNPS) [34] respectively into 500 Mass2Motifs each. These data sets are those used for training in [8]. In contrast to the beer data, none of these spectra are derived from Orbitrap instruments. The structural identity of all metabolites is known, providing a ground truth. In both cases, the Mass2Motifs characterized from beer were included in the analysis and kept fixed, while all other Mass2Motifs are learnt during LDA inference (details in SI Appendix Section S2.8). This therefore assesses the extent to which structurally characterized Mass2Motifs in one analysis can be used for metabolite annotations in another (from another instrument type). We manually verified all metabolites that include the formerly characterized Mass2Motifs and found that, at a probability threshold of 0.1, 81.5% and 63.3% of substructure annotations (for MassBank and GNPS, respectively) were validated (see S2.8 Fig. S-12 for detailed analysis of Mass2Motifs). In total, 694 (Massbank) and 613 (GNPS) spectra were found to have one or more validated substructure annotations (note that this is based solely on the Mass2Motifs annotated in the beer analysis, demonstrating a wide coverage from a small number of Mass2Motifs). MS2LDA also discovered MassBank and GNPS related substructures, complementary to those found in beer, showing its generic use. We repeated the analysis on a complex biological mixture (a human urine sample) and matched the Mass2Motifs discovered in beer to those found in urine. Matched standards in the urine are then used to validate the Mass2Motifs characterizations. At the 0.1 threshold, 74.3% of structural characterizations were validated. These results clearly demonstrate the potential of MS2LDA for substructure annotation.

One illustrative example of annotation with MS2LDA is provided in Figure 4. A subset of the network produced by MS2LDAvis (see SI Appendix Section S1.4) is shown consisting of molecules related to two Mass2Motifs (ferulic acid and ethylphenol). All but one molecule includes just one of the Mass2Motifs but one belongs to both (the fragments belonging to each Mass2Motif are clearly visible). The presence of both Mass2Motifs allows us to putatively annotate it as feruloyltyramine (314.1386 m/z; $[C_{18}H_{20}NO_4]^+$) despite spectral matching producing no relevant hits (see SI Appendix Table S-9). The output of Molecular Networking [19, 34] is shown on the right of Figure 4 (described in SI Appendix Section S2.9). This produces clusters

interpretable as ferulic acid and ethylphenol related, but as each molecule can belong to only one cluster, feruloyltyramine is assigned to the ethylphenol cluster and its relationship with ferulic acid is lost. Allowing each spectra to include multiple Mass2Motifs thus gives far greater potential in making de novo structural annotations of molecules. A lower perplexity of the LDA model compared to a standard multinominal model supports these results (SI Appendix Section S2.10). The phenomenon of individual spectra containing multiple correct substructure annotations is widespread. In the MassBank and GNPS datasets we counted the number of spectra associated with 1,2,3 and 4 different manually validated annotations from the beer characterized Mass2Motifs. Of the 694 Massbank spectra (613 GNPS) that had one or more validated substructure annotations, 212 (GNPS 34) had two or more, 39 (GNPS 4) three or more, and 3 four (GNPS 0) (see SI Appendix Fig. S-14).

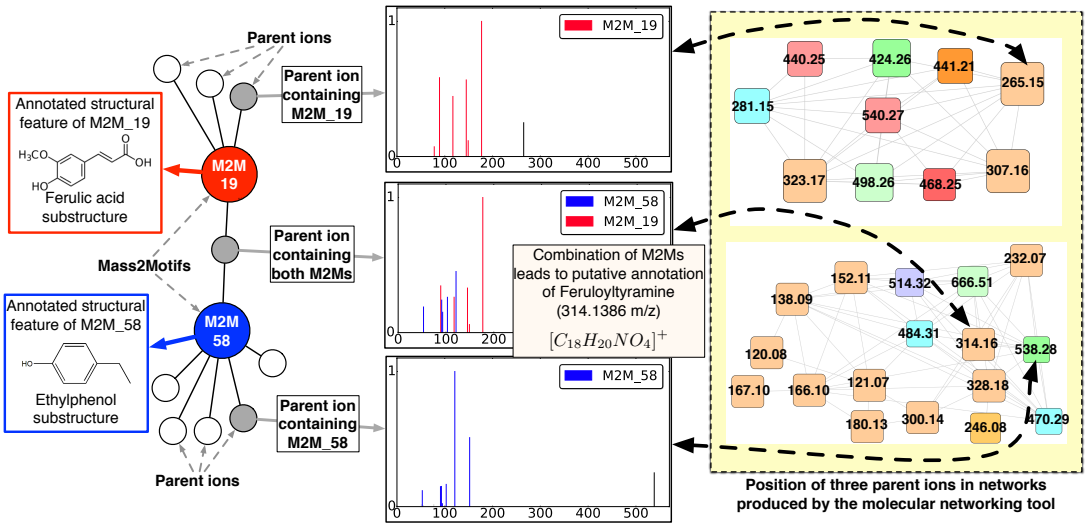


Figure 4: Mass2Motifs 19 and 58 were found to be representative of ferulic acid and ethylphenol, respectively. 11 and 42 MS1 features in the beer3 data set were explained by those two Mass2Motifs. Of those, one was explained by both, aiding in its annotation as feruloyltyramine (314.1386 m/z; $[C_{18}H_{20}NO_4]^+$). On the right of the plot, we show the clusters containing these MS1 features created using the molecular networking tool [34] (top: ferulic acid, bottom: tyramine (ethylphenol)). Node colouring and size are irrelevant here. The compound containing both Mass2Motifs is forced into the ethylphenol cluster, losing its relationship with ferulic acid.

2.4 Differential Expression of Mass2Motifs Reveals Biochemical Changes Across Samples

Annotating more metabolites is beneficial when investigating the changes in metabolite intensity across multiple samples. As MS2LDA groups metabolites in a biochemically relevant manner, we can go a step further and consider the differential expression of Mass2Motifs in a manner similar to approaches taken in transcriptomics where it is common to consider the shared differential expression (DE) of a group of related transcripts as indicative of their contribution to a common aspect of cellular biology [28]. For example, consider a standard metabolomics experiment comparing MS1 intensities across multiple replicates of two conditions. After the MS1 peaks have been matched across samples, those that share a Mass2Motif (defined in a single MS2LDA analysis of one of the samples or an additional pooled sample) can be grouped, and the DE of the groups computed. To demonstrate, we compared three full-scan replicates of beers 2 and 3 using MS1 groupings defined by the Mass2Motifs from the MS2LDA analysis of Beer3. DE of groups was assessed using PLAGE [29]. Figure 5A shows MS1 peaks assoociated with a gunanine-related

Mass2Motif suggesting that in Beer3 free guanine is more abundant whereas in Beer2, guanine-conjugates dominate. Similarly, molecules associated with the pentose Mass2Motif (Figure 5B) show DE between beers 2 and 3. We investigated whether or not similar outcomes could be achieved with spectral similarity clustering. However, the 12 pentose-related metabolites were distributed across 10 clusters hiding the correlated intensity change (see SI Appendix Section S2.11 for more examples).

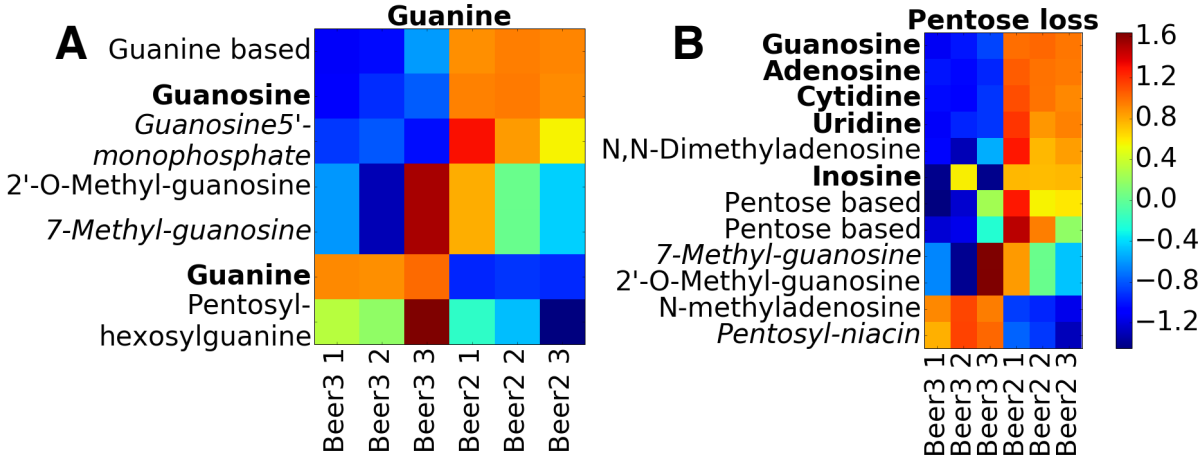


Figure 5: Log fold change heat-maps for the **A)** guanine and **B)** pentose loss Mass2Motifs. Each row is an MS1 peak and columns represent samples. Bold names could be matched to a reference compound. Detailed annotations of metabolites can be found in SI Appendix Table S-18.

3 Discussion

MS2LDA was inspired by the idea that conserved fragments and neutral losses can be indicative of metabolite substructures and the implied parallel with topic modelling of text. No alternative tools exist that allow for the unsupervised substructure mining from MS fragmentation data whilst also allowing for multiple such substructures to be present within one metabolite. MS2LDA can group molecules that share substructures without high similarity across their entire MS2 spectra. It reduces complex fragmentation data sets into metabolites explained by one or more patterns of concurring mass fragments or neutral losses – Mass2Motifs.

MS2LDA relies on reliable matching of MS1 peaks to MS2 spectra and works best for complex mixtures where a large number of metabolites are fragmented and information-rich MS2 spectra are available (e.g. generated by ramped or stepped collision energy). High-resolution MS fragmentation can differentiate mass fragments and neutral losses even at low mass range of 50-70 m/z (see SI Appendix Section S2.12). Manual structural characterization of many Mass2Motifs is straightforward and the structural features or substructures can be propagated to all connected MS2 spectra. Based on initial experiments, automated Mass2Motif annotation is promising (19 of the characterized positive mode beer Mass2Motifs were correctly annotated, despite the fact that losses are not currently supported by spectral matching tools and had to be omitted, see SI Appendix Section S2.13).

Metabolite annotation and identification is a bottleneck in high-throughput metabolomics. MS2-LDA can assist by automatically assigning possible substructures to a fragmented LC-MS peak via the Mass2Motifs present in its MS2 spectrum. MS2LDA can thus quickly classify MS1

peaks into functional classes without knowing the complete structure of the metabolite. On average, over 70% of the fragmented metabolites were explained by one or more structurally annotated Mass2Motifs, a massive improvement on results reported in a recent study, again using beer as an exemplar, where only 2-3% of the high-abundance differentially expressed molecular features could be classified [1]. Validation on data from the MassBank and GNPS databases also demonstrated the validity of our structurally characterized Mass2Motifs and also showed how fixed Mass2Motifs characterized in one analysis could be used in other data sets, even those produced from different labs on different instruments. In addition, the biochemically relevant metabolite grouping provided by MS2LDA allows us to identify Mass2Motifs that are enriched with metabolites with correlated intensity variation.

Computationally, MS2LDA is more costly than simpler tools, but not prohibitively so. For example, using Variational Bayesian inference, the GNPS data set (5670 spectra) could be decomposed into 500 Mass2Motifs in approximately 4 hours on a laptop. As LDA has been used on very large text corpora (e.g. 3.3 million documents from Wikipedia [12]), the technology exists to comfortably scale this type of analysis to larger metabolomic data sets. In addition, we envisage MS2LDA being used in conjunction with a standard MS1 analysis via fragmentation of a pooled sample from which Mass2Motifs can be linked to MS1 intensity variability as described in the differential expression section above.

The MS2LDA approach is markedly different from other analysis tools as multiple Mass2Motifs can be associated with one metabolite, and determination of the fragments / neutral losses that are part of a conserved structural motif is unsupervised. Our proposed focus on mining the MS2 fragmentation data alone to aid in identification of functional classes of metabolites is unique and complementary to existing use of fragmentation data. We anticipate MS2LDA to be particularly useful in research areas such as clinical/pharmaco and nutritional metabolomics, environmental analysis, and natural products research, as it can quickly recognize substructure patterns related to drugs and food-derived metabolites in an unsupervised way. Although we have demonstrated MS2LDA on DDA data, we see no reason why it would not work on data independent acquisition (DIA) data in which fragments have been matched to MS1 ions using, e.g. MS-DIAL [30].

4 Materials & Methods

All data and code are available from http://dx.doi.org/10.5525/gla.researchdata.313.

4.1 Materials

Four beer samples were used as representative of diverse complex mixtures (see SI Appendix Section S3). 10 ml of beer was sampled directly after opening and stored at -20°C before extraction. After thawing, i) 200μL of beer was mixed with 600μL of methanol/chloroform, ii) sonicated for 5 minutes at room temperature; iii) and centrifuged for 5 minutes (12,000 g) at room temperature. The supernatants were stored at -80°C. Urine fragmentation data from an earlier approved and published study on metabolite annotation of urinary metabolites was used for validation purposes [32]. HPLC-grade methanol, acetonitrile, and analytical reagent grade chloroform were acquired from Fisher Scientific, Loughborough, UK. HPLC grade H2O was purchased from VWR Chemicals, Fountenay-sous-Bois, France. Formic acid (for MS) and ammonium carbonate were

acquired from Fluka Analytical (Sigma Aldrich), Steinheim, Germany.

4.2 Methods

A Thermo Scientific Ultimate 3000 RSLCnano liquid chromatography system (Thermo Scientific, CA, USA) was coupled to a Thermo Scientific Q-Exactive Orbitrap mass spectrometer equipped with a HESI II interface (Thermo Scientific, Hemel Hempstead, UK). Thermo Xcalibur Tune software (v2.5) was used for instrument control and data acquisition. Column temperature was maintained at 25 °C. The hydrophilic interaction liquid chromatography separation was performed with a SeQuant ZIC-pHILIC column (150 x 4.6 mm, 5 μm) equipped with the corresponding pre-column (Merck Sequant, Darmstadt, Germany). A linear LC gradient was conducted from 80% B to 20% B over 15 min, followed by a 2 min wash with 5% B, and 7 min re-equilibration with 80% B, where solvent B is acetonitrile and solvent A is 20 mM ammonium carbonate in water. The flow rate was 300 μL/min, column temperature held at 25 °C, injection volume was 10 μL, and samples maintained at 4 °C in the autosampler [31]. Samples were measured in randomized order [6] (see SI Appendix Section S4). MS and MS/MS settings can be found in SI Appendix Section S5. For positive and negative ionization combined fragmentation mode, the duty cycles consisted of a full scan in positive ionization mode, followed by a TopN data dependent MS/MS (MS2) fragmentation event taking the 10 most abundant ion species not on the dynamic exclusion list, followed by the same two scan events in negative mode. MS/MS fragmentation spectra were acquired using stepped higher collision dissociation (HCD) combining 25.2, 60.0, and 94.8 normalized collision energies (NCEs) in one MS2 scan. In full scan mode, the duty cycle consisted of two full scan events. The duty cycles for positive and negative ionization separate fragmentation modes, respectively, consisted of one full scan (MS1) event and one Top10 MS/MS (MS2) fragmentation event.

5 Acknowledgement

We thank Dr Emma Schymanski (RMassBank), Dr Tony Larson (xcmsFragments), and Dr Samuel Bertrandt (the seven golden rules) for assistance with implementation of the mentioned R scripts; Kai Duhrkop for providing us the GNPS and MassBank spectra used in the CSI:FingerID paper, and Dr Niels van den Broek for helpful discussions on acquisition of fragmentation spectra. JJJvdH was supported by the Wellcome Trust [105614/Z/14/Z]. MPB was funded by the Wellcome Trust Centre for Molecular Parasitology [104111/Z/14/Z]. JW was supported by a Scottish Informatics and Computer Science Alliance PhD studentship. SR was supported by Biotechnology and Biological Sciences Research Council BB/L018616/1.

References

[1] F. Allen, R. Greiner, and D. Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, 2015.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] S. Böcker and Z. Lipták. A fast and simple algorithm for the money changing problem. *Algorithmica*, 48(4):413–432, 2007.

[4] S. Böcker, M. C. Letzel, Z. Lipták, and A. Pervukhin. Sirius: decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[5] X. Chen, X. Hu, X. Shen, and G. Rosen. Probabilistic topic modeling for genomic data interpretation. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 149–152. IEEE,Piscataway,NJ, 2010.

[6] D. J. Creek, A. Jankevics, R. Breitling, D. G. Watson, M. P. Barrett, and K. E. V. Burgess. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography–mass spectrometry: improved metabolite identification by retention time prediction. *Analytical Chemistry*, 83(22):8703–8710, 2011.

[7] R. R. da Silva, P. C. Dorrestein, and R. A. Quinn. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112(41):12549–12550, 2015.

[8] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.

[9] W. B. Dunn, A. Erban, R. J. M. Weber, D. J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9(1):44–66, 2013.

[10] N. Garg, C. A. Kapono, Y. W. Lim, N. Koyama, M. J. A. Vermeij, D. Conrad, F. Rohwer, and P. C. Dorrestein. Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures. *International Journal of Mass Spectrometry*, 377:719–727, 2015.

[11] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[12] M. Hoffman, D. Blei, and F. Bach. Online learning for latent dirichlet allocation. *Advances in Neural Informatics Processing Systems 23*, pages 1–9, 2010. ISSN 08912017. doi: 10.1.1.187.1883.

[13] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010.

[14] F. Hufsky, K. Scheubert, and S. Böcker. Computational mass spectrometry for small-molecule fragmentation. *Trends in Analytical Chemistry*, 53:41–48, 2014.

[15] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7(1):1–10, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-234. URL `http://dx.doi.org/10.1186/1471-2105-7-234`.

[16] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8(1):art. no. 105, 2007.

[17] Y. Ma, T. Kind, D. Yang, C. Leon, and O. Fiehn. Ms2analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Analytical Chemistry*, 86(21):10724–10731, 2014.

[18] B. B. Misra and J. J. J. der Hooft. Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis*, 37(1):86–110, 2016.

[19] D. D. Nguyen, C.-H. Wu, W. J. Moree, A. Lamsa, M. H. Medema, X. Zhao, R. G. Gavilan, M. Aparicio, L. Atencio, C. Jackson, et al. MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences*, 110(28):E2611–E2620, 2013.

[20] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. J. Bino, and J. Vervoort. Automatic chemical structure annotation of an LC–MSn based metabolic profile from green tea. *Analytical Chemistry*, 85(12):6033–6040, 2013.

[21] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, J. Vervoort, and R. J. Bino. In silico prediction and automatic LC–MSn annotation of green tea metabolites in urine. *Analytical Chemistry*, 86(10):4767–4774, 2014.

[22] S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cdna microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(2):143–156, 2005.

[23] R. A. Scheltema, A. Jankevics, R. C. Jansen, M. A. Swertz, and R. Breitling. Peakml/mzmatch: a file format, java library, r library, and tool-chain for mass spectrometry data analysis. *Analytical Chemistry*, 83(7):2786–2793, 2011.

[24] C. Sievert and K. E. Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[25] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.

[26] M. A. Stravs, E. L. Schymanski, H. P. Singer, and J. Hollender. Automatic recalibration and processing of tandem mass spectra using formula annotation. *Journal of Mass Spectrometry*, 48(1):89–99, 2013.

[27] D. L. Sweeney. A data structure for rapid mass spectral searching. *Mass Spectrometry*, 3 (Special_Issue_2):S0035–S0035, 2014.

[28] A. L. Tarca, G. Bhatti, and R. Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS One*, 8(11):e79217, 2013.

[29] J. Tomfohr, J. Lu, and T. B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):art. no. 225, 2005.

[30] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, and M. Arita. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Meth*, 12(6):523–526, jun 2015. ISSN 1548-7091.

[31] J. J. J. van der Hooft, R. C. H. de Vos, L. Ridder, J. Vervoort, and R. J. Bino. Structural elucidation of low abundant metabolites in complex sample matrices. *Metabolomics*, 9(5): 1009–1018, 2013.

[32] J. J. J. van der Hooft, S. Padmanabhan, K. E. V. Burgess, and M. P. Barrett. Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics*, 12(7):1–15, 2016.

[33] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, et al. Hmdb 3.0—the human metabolome database in 2013. *Nucleic Acids Research*, page gks1065, 2012.

[34] J. Y. Yang, L. M. Sanchez, C. M. Rath, X. Liu, P. D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. de Felicio, A. Fenner, et al. Molecular Networking as a dereplication strategy. *Journal of Natural Products*, 76(9):1686–1699, 2013.

Topic Modeling for Untargeted Substructure Exploration in Metabolomics
van der Hooft et al.

Supporting Appendix

TABLE OF CONTENTS

SECTION S1. MS2LDA WORKFLOW

This section described the entire MS2LDA workflow developed within this study.

The entire MS2LDA workflow is summarized in Figure S-1.

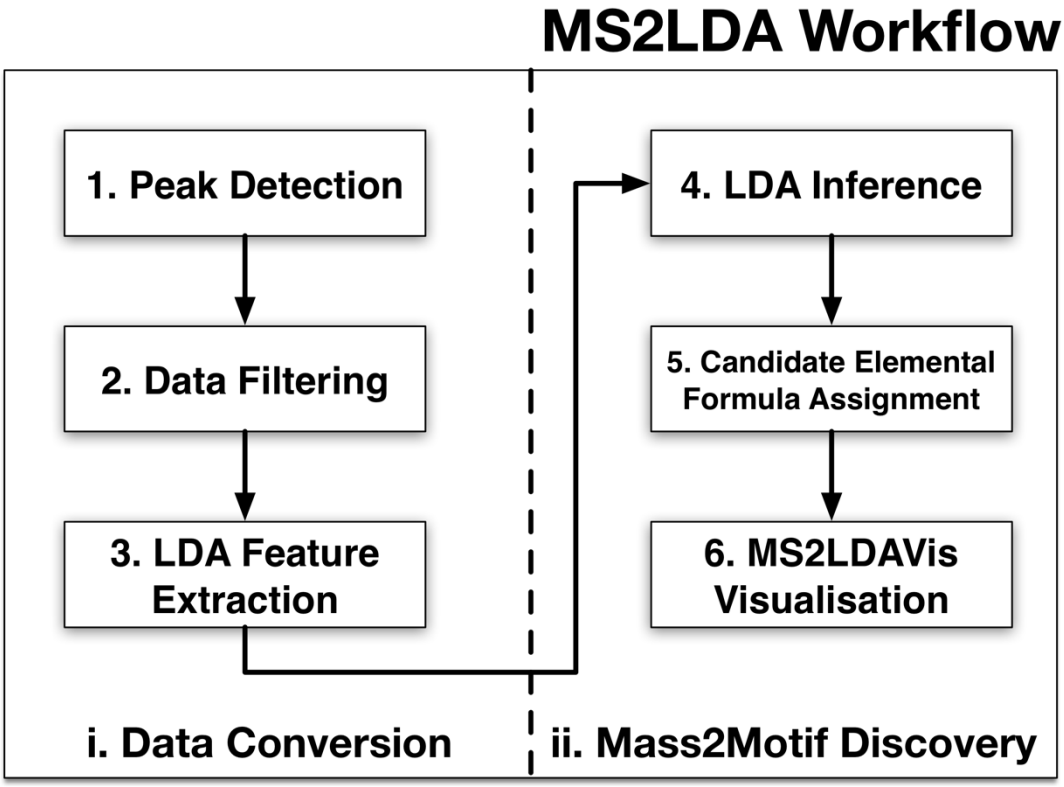


Figure S-1. The MS2LDA workflow.

S1.1 Data Conversion Stage

Data conversion is an essential part of the MS2LDA workflow, since the acquired fragmentation data cannot readily be used for the purpose of mass fragmental pattern searching. Our workflow (illustrated in Figure S-1) accepts as input the combination of a single full-scan file for the MS1 peaks and a separate fragmentation file for the MS2 peaks (alternative strategies for peak detection and MS1-MS2 correspondence establishment that accept different combinations of input files, such as using just a single fragmentation file for both the MS1 and MS2 peaks, are also provided in our workflow). The data conversion process starts with the detection of MS1 peak in the input .mzXML file obtained from full-scan mode spectra using the CentWave algorithm from XCMS (1) and the .mzML file obtained in MS/MS mode. Matching of a parent (MS1) LC-MS peak to fragment (MS2) peaks are then established using a script based on the RMassBank package (2), through greedy search for the most intense unique MS2 spectrum (more intense fragmentation spectra are generally information-richer) that can be linked to an MS1 LC-MS peak within a specified retention time (RT) window. A filtering step based on RT and intensity is applied to remove noisy peaks, as well as the washing part, equilibration part, and the start of the chromatogram prior to the injection peak. Finally, any MS1 peak not having paired MS2 peaks is discarded. This process leaves unique MS1-MS2 pairs, thereby omitting the lower intense fragmentation spectra of MS1 peaks that were fragmented multiple times. This greatly helps in the LDA modelling, as multiple spectra of the same MS1 peak could be considered as conserved mass fragmental motif in the data set.

MS1 feature extracted →

Binned MS2 fragments and losses ↓

	MS1_a	**MS1_b**	**MS1_c**	**MS1_d**	**MS1_e**	**…**
Fragment_119.0351	0	100	24	37	0	
Fragment_136.0629	87	0	17	18	0	
Fragment_156.0769	55	20	0	10	100	
…						
Loss_18.0080	56	0	0	10	15	
Loss_36.0183	0	0	30	0	0	
Loss_46.0053	40	40	10	87	100	
…						

Figure S-2. The data-frame extracted from fragmentation data: a matrix of XCMS-picked MS1 peaks (columns) and binned mass fragment (and neutral loss) features with normalized (0 – 100 scale) intensities.

The next step in the data conversion stage is the transformation of the spectral data into a format that is suitable for concurring pattern discovery, which is a matrix consisting of the MS1 peaks (columns) and their correspondent MS2 fragments (rows) or losses (see Figure S-2). Drawing an analogy from text processing, each MS1 peak can now be seen as a 'document' while the linked MS2 spectrum associated to each MS1 peak produce the 'word' features in a document. Note that following the bag-of-words assumption, LDA does not take into account the word order, but merely the word count, i.e., the number of times a word occurs in a document (further described in Section S1.2). For each MS1 peak, two types of word features can be extracted from a MS2 fragmentation spectrum:

- **Fragment features**, which are the discretized mass values of the MS2 peaks. A greedy binning process is used to group MS2 peaks within a certain user-defined m/z window from the next unprocessed MS2 peak. This way, MS2 peaks with close-enough m/z values but observed in different precursor MS1 peaks are linked and placed into the same discrete bin – each bin corresponds to a fragment feature. The input for inference in textual LDA is the count of occurrences of words in each document; in MS2LDA, the intensity values of MS2 peaks can be considered to be proxies for word counts. These intensity values are normalized by dividing to the largest intensity value in the fragmentation spectrum and discretized on a scale of 0 to 100 (integers).

- **Loss features**, which is the discretized mass values of the neutral losses. Neutral losses are the mass differences between a precursor MS1 peak and each of its MS2 peaks in the spectrum. To produce the loss features, we find the m/z difference between each fragment peak to its precursor ion. Similar to fragment features, the normalized intensity values of the neutral losses, represented by the intensities of their resulting mass fragments, are used as proxies for the loss counts.

We provide a single configuration file that encodes user-defined data conversion parameters as key-value pairs in a human-readable textual (YAML) format, as many parameters of the data conversion step are platform-dependent. A single 'R' script is then provided to load all the user-defined parameters from the configuration file and start the entire data conversion step. An example configuration file, used for data processing and analysis in this paper, is provided in our online repository.

Special feature extraction pipeline for GNPS and MassBank

For validations of fixed Mass2Motifs (learnt from the beer dataset) that were applied to the GNPS and MassBank datasets, an alternative feature extraction pipeline was required. Firstly, a parser was written to read GNPS and MassBank datasets that are available in the .MGF format. Gaussian kernel density estimation was used to combine fragments and neutral losses observed in different spectra into a global fragment vocabulary. This was found, via visual inspection, to produce better fragment groupings for this data than the mass binning approach in Section S1.1. Gaussian kernel widths were set such that 3 standard deviations were equal to 7ppm for the fragments features and 15ppm for the loss features (the higher value for the loss features is justified by the fact that they are computed as the difference between two noise measurements). Features are extracted as the modes (maxima) of the density estimate with their width determined by when the density hits a minimum or the width exceeds a maximum (50ppm).

S1.2 Mass2Motif Discovery Stage

Given the matrix of features co-occurrences produced from the data conversion stage, our goal is to infer the concurring patterns of features shared by the fragmentation spectra. Following the Latent Dirichlet Allocation (LDA) model, a fragmentation spectrum can be seen as a mixture over potentially substructure patterns (which we called Mass2Motifs), each of which is itself a distribution over fragment/loss word features. A fragmentation spectrum, linked to a particular MS1 peak, can therefore be generated in this model by firstly sampling for the Mass2Motifs that the spectrum is comprised of and subsequently sampling the specific fragment/loss features from the selected Mass2Motifs. A brief summary of the LDA model in the context of fragmentation data and the inferential procedure is described next. To infer the latent Mass2Motifs present in the data, a Python implementation of a collapsed Gibbs sampling scheme is used in our MS2LDA workflow (3).

We assume the bag-of-word word model, where within each fragmentation spectrum the observed MS2 word features are exchangeable, i.e., their order does not matter, only their observed counts (intensities) matter. Given some K Mass2Motifs (indexed by $k = 1, \dots, K$), the observation of the n-th word in the d-th MS1 document can be described by the following generative process:

$$w_{dn} | \varphi_{z_{dn}} \sim Multinomial(\varphi_{z_{dn}})$$

$$z_{dn} | \theta_d \sim Multinomial(\theta_d)$$

$$\theta_d | \alpha \sim Dirichlet(\alpha)$$

$$\varphi_k | \beta \sim Dirichlet(\beta)$$

In other words, observation on the n-th word in the d-th MS1 fragmentation spectra (w_{dn}) is conditioned on the assignment of MS2 fragment/loss word w_{dn} to some k-th Mass2Motif multinomial distribution (corresponding to a concurring pattern of fragments and/or losses). This assignment is denoted by the indicator variable z_{dn}, so $z_{dn} = k$ if w_{dn} is assigned to a k-th multinomial. The k-th multinomial distribution that an MS2 word is assigned to is characterized by the parameter vector $\varphi_{z_{dn}}$. However, $\varphi_{z_{dn}}$ is itself drawn from a prior Dirichlet distribution with parameter vector β. The probability of seeing certain Mass2Motifs for each d-th fragmentation spectra is then drawn from a multinomial distribution with a parameter vector θ_d. This parameter vector θ_d is in turn drawn from a prior Dirichlet distribution having parameter vector α. Intuitively, if we assume symmetric prior on the α and β vectors (i.e. they are scalar), a high value set on α means each fragmentation spectra is likely to contain a mixture of most Mass2Motifs, while lower values on α means fragmentation spectra will contain fewer Mass2Motifs. Similarly, higher β means a Mass2Motif is likely contain a mixture of most words, while lower β leads to a Mass2Motif containing a mixture of fewer words.

Given the matrix of fragment/loss word counts produced from the feature extraction step and user-defined choices of hyper-parameters (α, β, K) that suit the input data, the posterior distributions of documents-to-topics (all the θ_ds) and topics-to-words (all the φ_ks) can be approximated.

Gibbs sampling

We follows the method described by (3) and uses a collapsed Gibbs sampling scheme to perform inference. Gibbs sampling is an instance of Markov chain Monte Carlo algorithm commonly used to approximate posterior distributions in Bayesian inference where direct sampling or closed form solutions are difficult to obtain. In this particular case of LDA inference, the input to Gibbs sampling is the observed counts of fragment/loss words co-occurrences in fragmentation spectra (documents) and as output, we infer the latent Mass2Motif-to-words distributions and fragmentation spectra-to-Mass2Motif distributions present in the data.

Since Dirichlet priors are conjugate to the multinomial distributions θ and φ, we can marginalize out the θ and φ parameters. Assuming a symmetric prior probability distribution on α and β, the conditional probability for the assignment of the n-th fragment/loss word feature in the d-th fragmentation spectrum (linked to a particular MS1 peak) to the k-th Mass2Motif is denoted here:

$$P(z_{dn} = k | w_{dn}, \dots) \propto \frac{c_{kn} + \beta}{c_k + N\beta} \cdot c_{dk} + \alpha$$

where:

- c_{kn} is the count of the number of word n in the vocabulary that are currently assigned Mass2Motif k
- c_k is the count of all words currently assigned to Mass2Motif k
- c_{dk} is the count of words from MS1 peak d assigned to Mass2Motif k

All these counts are computed after removing the current word w_{dn} being iterated upon in the Gibbs sampling step. Finally, to approximate the document-to-topic distributions (θ_d for each MS1 peak or document d) and the topic-to-word (or Mass2Motif to fragment or loss feature) distributions (φ_k for each topic k), we use the expectation of a Dirichlet distribution, the expected values of the parameters θ and φ given w and z are:

$$\theta_{dk} = \frac{c_{dk} + \alpha}{c_d + K\alpha}$$

$$\varphi_{kn} = \frac{c_{kn} + \beta}{c_k + N\beta}$$

In our Gibbs sampling implementation, only the last sample (after monitoring for convergence) was used for the purpose of analysis (as an alternative, we can also average the posterior estimates over the samples, although we found no discernible difference between using the final sample and using the mean taken over multiple samples). Due to the stochastic nature of the Gibbs sampling procedure, we might get slightly different results each time, which may be undesirable. To overcome this, we set a constant random seed for the sampler, allowing us to get the same inference results each time, provided the same parameters of K, α, β are used with the same input files.

Variational inference

In addition to Gibbs sampling, we have also implemented Variational Bayesian inference for LDA using the algorithm described in (5). In essence, the variational method approximates the intractable posterior density via a product of densities which are updated in an iterative manner until convergence. Once converged, the algorithm provides the Mass2Motif to feature distributions, as well as Dirichlet distributions for the spectra to Mass2Motif relationship and the global Mass2Motif relationship. In our experiments we have found no discernible difference between the output of the Gibbs sampler and Variational Bayesian implementations although the Variational Bayes method is faster (see the Running Times section below).

Cross-validation

The number of Mass2Motifs and model fit are estimated via a 4-folds cross-validation approach. For each test fold being held out in the fragmentation spectra data set, an estimate of the model evidence is computed after training the model on the remaining training folds in the data set. A comparison of LDA against the multinomial

mixture model (clustering) is provided in Section S2.10. A crucial difference between LDA and standard mixture-model clustering lies in the modelling assumption that a document is a mixture of one or more topics (LDA) as opposed to each document having exactly one topic (clustering). We compare the model fit of LDA against clustering by evaluating the log evidence and perplexity on a held-out beer data file (beer3 positive ionization mode). The perplexity measures how well a probability distribution or probability model predicts a sample and is defined as:

$$perplexity(W) = exp\left(\frac{\sum_d \log(P(w_d))}{\sum_d N_d}\right)$$

where $perplexity(W)$ is the perplexity on the whole held-out test collection, $P(w_d)$ is the marginal probability of a testing document d (integrating over all the parameters of the model), approximated via an importance sampling method as described by Wallach et al. (4) and N_d is the number of words in each testing document d. We follow (3) and set the value of the hyperparameters α =K/50 and β =0.1 for LDA during the cross-validation experiment. For mixture model clustering, a non-informative Dirichlet prior (with constant parameter α =K/50, where K is now the number of clusters) is set on the proportions of the mixture components and another Dirichlet prior (with constant hyper-parameter β =0.1) is set on cluster-specific word distributions. The Gibbs sampler for LDA and multinomial mixture model is run for 1000 samples, discarding the first 500 for burn-in. The lower perplexity (shown in Section 2.10, Figure S-15) demonstrates that LDA provides a better model fit on the held-out data compared to multinomial mixture model.

Incorporating previously defined Mass2Motifs

In our experiments on Massbank, GNPS and urine data, we incorporated Mass2Motifs from the beer analysis into the MS2LDA framework. This is straightforward within the Variational Bayesian framework if features can be matched across the two analysis. In particular, when updating the Mass2Motif to feature probability distributions, we can leave some (the previously defined ones) unchanged and just update the others – i.e. our model consists of static, previously defined Mass2Motifs and new, learnable ones. In our experiment, we fixed the ~30 Mass2Motifs that where characterized in beer and updated the other 470 in the Variational Bayesian inference routing. To match the features, we took the features present in each of the characterized beer Mass2Motifs and searched for them in the features generated for the new analysis. For each Mass2Motif, we added up the feature probabilities for those that could be matched. A Mass2Motif was included in the new analysis if features making up at least 0.5 of their probability could be matched.

Running times

We provide an illustrative example of the running time of the MS2LDA pipeline for a beer sample on a laptop (Intel Core i7, 16GB RAM). The data conversion stage includes the peak detection step via the CentWave algorithm from XCMS, the linking of parent (MS1) peak to fragment (MS2) peaks using the script based on RMassBank, as well as the binning process to create fragment and loss features. This was completed in 20 minutes and produces a matrix of features co-occurences that can be used for LDA inference. During inference, running Gibbs sampling with 1000 posterior samples requires approximately an hour. The alternative of running Variational Bayesian inference with 1000 steps takes half an hour.

The running time required for the data processing and inference steps of a single sample in MS2LDA is therefore approximately 1.5 hours in total.

S1.3 Candidate Elemental Formula Assignment

The MS2LDA workflow provides two optional methods to assign candidate elemental formulae to the mass fragments, neutral losses, and precursor ions. The first is achieved by integrating SIRIUS (**S**um formula **I**dentification by **R**anking **I**sotope patterns **U**sing mass **S**pectrometry, (6)) into our workflow. SIRIUS assigns elemental formula by posing it as an integer decomposition problem and solving it through a dynamic

programming approach ('Round Robin') (7). SIRIUS is freely-available and, as it is written in Java, can in theory be run platform-independently on any Windows, Unix and Mac environment (in practice, library dependencies have to be satisfied before SIRIUS can be run on the target computer). Integration of SIRIUS into our workflow is achieved by wrapping calls to the Java package of SIRIUS through a separate sub-process, passing it a temporary .MGF file that corresponds to each fragmentation spectrum. SIRIUS assigns elemental formulae to each combination of MS1 and MS2 peaks independently, which may lead to mass fragments of similar m/z value being assigned an elemental formula in some spectra, but not in all.

As an alternative strategy for annotation, our workflow also provides a pure Python implementation of an elemental formula assigner (called 'EF-Assigner') based on the Round Robin algorithm that also lies at the heart of SIRIUS. Once the initial assignment of potential candidate formulae to mass fragments, neutral losses and also precursor ion masses has been performed, the list of candidate formulae is further filtered using our implementation of the 7-golden rules, a set of heuristic rules introduced by Kind and Fiehn (8). This filtering step is used to remove chemically-unlikely elemental formula compositions from the candidate list. Advantages of the EF-Assigner module are its easy compatibility to the MS2LDAvis module (which is also written in Python) and it assigns elemental formulae to the binned fragments and losses in the matrix instead of to individual spectra. However, unlike SIRIUS that uses the complete information of the precursor ion and fragments peaks in a spectrum for annotation, EF-Assigner assigns the elemental formulae for the MS1 peaks, mass fragments and neutral losses independently.

S1.4 Visualisation Using the MS2LDAvis Module

Inference results from LDA can be challenging to interpret due to the (still) high dimensionality of the data. Analysis of Mass2Motifs to examine if they correspond to actual structural features or biochemical substructures is an iterative and exploratory process. In our workflow, this is made possible through the MS2LDAvis module -- an interactive web-based visualization that can be used to explore and validate Mass2Motifs from MS2 data. MS2LDAvis is extended from the Python port of the topic modelling visualization interface LDAvis (9), which is built upon the combination of the Javascript/D3 library. While initially based on LDAvis, the MS2LDAvis module has been greatly customized to suit our Mass2Motifs and fragmentation data exploration needs.

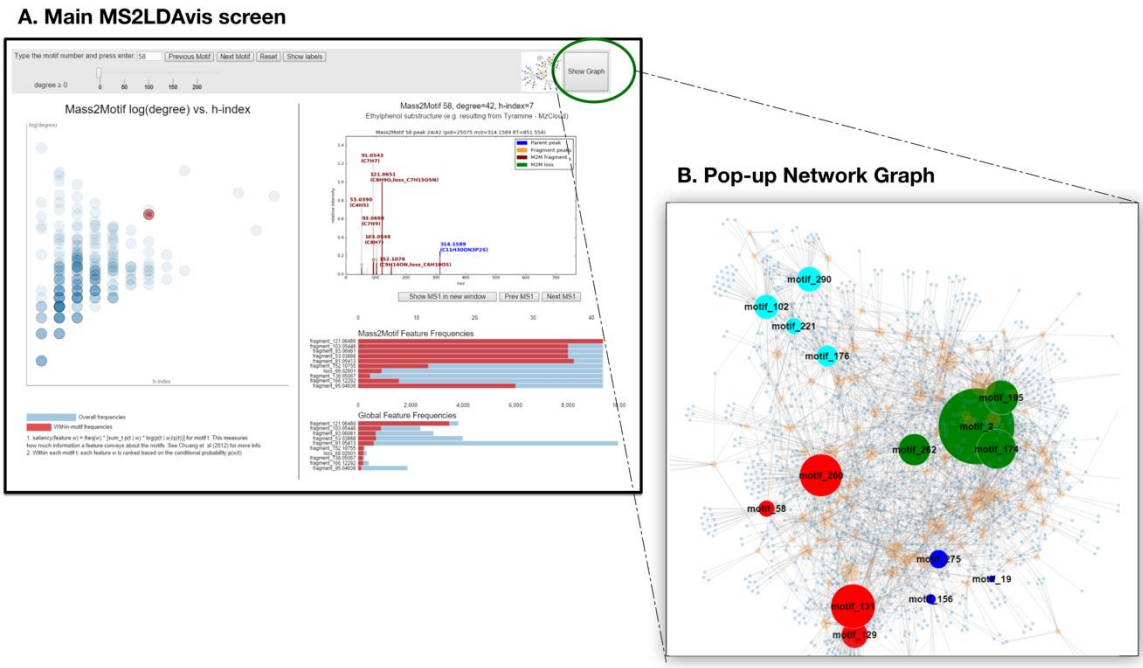


Figure S-3. A) The main MS2LDAvis screen, while B) is the network graph of beer3 extract positive ionization mode file where a number of Mass2Motifs were selectively colored before loading the network visualization. Mass2Motifs circles are proportional to their degree (number of connections), whereas small blue squares represent fragmented MS1 peaks.

Similar to the original LDAvis, the left panel of our MS2LDAvis module shows a global view of the model, whilst the right panel zooms into a specific Mass2Motif (see Figure S-3A). However, unlike LDAvis where topics are displayed on the left panel through multidimensional scaling that projects topics to two dimensions, the two axes in our MS2LDAVis panel are the log-degree and the h-index of Mass2Motifs. We defined the *degree* of a Mass2Motif as the number of fragmentation spectra explained by the Mass2Motif at the user-defined thresholding level t_θ on the fragmentation-spectra-to-Mass2Motif distributions (the θ parameters). The h-index of a Mass2Motif is defined in a similar manner to the conventional h-index for scientific publications of a researcher. A Mass2Motif has an index of h if it has h fragment/loss features obtained after setting a user-defined threshold t_φ on the Mass2Motif-to-word distributions (the φ parameters), each of which occur in the set of thresholded documents at least h times. Intuitively, Mass2Motif with high degrees but low h-index could potentially correspond to simple structural features or substructures that occur in many MS2 fragmentation spectra, while Mass2Motif with high h-index but lower degrees could potentially correspond to more unique and complex substructures shared by fewer MS2 spectra.

The left and right panels of our visualization are linked such that selecting a Mass2Motif on the left changes the information displayed on the right panel. We further enhanced MS2LDAvis by plotting the fragmentation spectra of each MS1 peak (documents) above the user-defined threshold t_θ in the selected Mass2Motif. The fragment and loss words in the fragmentation spectra that are explained by the currently selected Mass2Motif, i.e., above the user-defined threshold t_φ, are highlighted in bold and user can easily flip through different fragmentation spectra explained by the topic by clicking the *Previous MS1* and *Next MS1* buttons under the fragmentation spectra plot. The bottom of the right panel displays two feature frequency histograms; the Mass2Motif Feature Frequencies histogram displays the counts of each Mass2Motif associated fragment or loss (above the user-defined threshold t_φ on the Mass2Motif-to-features distributions [the φ parameters]) within the fragmentation spectra explained by the Mass2Motif. Similarly, the Global Feature Frequencies histogram displays the overall frequency of the fragments or losses within the complete data set that can be explained by the currently selected Mass2Motif. This provides an estimate of how unique the fragment/loss features are in the whole data set.

Finally, to complement our main view, we also allow the possibility of exploring the inferred substructure data in a pop-up network graph (see Figure S-3B), where Mass2Motifs and MS1 peaks form the nodes in the graph and edges are drawn between them if a document is explained by a topic with conditional probability above the user-defined threshold t_θ. The graph view can be accessed by clicking on the *Show Graph* button on the top panel of the main window. To minimize clutter in the network graph, user can also define a threshold on the degree of the Mass2Motifs, i.e., all Mass2Motifs with a degree of 10 or lower can easily be removed from the graph. Nodes in the graph can also be shown, hidden and coloured according to user-defined specifications before the visualisation interface is called (see Figure S-2B). The two complementary views are linked such that clicking a topic node on the network graph will select the corresponding topic on the main view and vice versa. The network graph is particularly useful in exploring the relationships between Mass2Motifs and investigating which MS1 peaks have fragmentation spectra that can be explained by multiple Mass2Motifs.

SECTION S2. SUPPORTING RESULTS

This section contains all the supporting Figures and Tables and accompanying explanatory texts that support the results in the manuscript. Please note that Supporting Tables S-4 and S-5 can be found as separate Word files.

S2.1 Mass2Motif structural characterizations

All Mass2Motifs in positive and negative mode ionization files with degrees of 10 or more were investigated to see if they represented any biochemical relevant substructure or structural feature. The resulting structural characterizations were collected in tabular format. Information on the key mass fragments, neutral losses, and degrees across the four beer files is shown in the Tables. A confidence was given to the structural characterization based on the collected evidence, using spectra matching to MzCloud (www.mzcloud.org) and expert knowledge.

Table S-4. Table with Mass2Motifs (MSMs) discover in the four positive ionization mode fragmentation files of the beer extracts.

Table with M2Ms in four beers – fragments/losses associated to the M2Ms can slightly differ in between beers due to degree and type of metabolites associated to the M2M. Experimental masses are within 5 ppm of theoretical masses as found in topics. Slight changes are observed per file. Annotated Mass2Motifs: Bold represents highest level of confidence (i.e., several fragments or specific mass value that can only point to a certain combination of ions), bold and italic is second-highest level of confidence (i.e., match on elemental formula (EF) only – but in the given sample matrix it is quite a likely structural annotation), just italic is the third-highest level of confidence (i.e., no specific structure found, often generic fragments that have multiple possible structural confirmations), and plain text represents the lowest level of confidence.

Beer1		**Beer2**		**Beer3**		**Beer4**		**Frag/ Loss**	**m/z**	**EF**	**Characterization**
M2M	**Degree**	**M2M**	**Degree**	**M2M**	**Degree**	**M2M**	**Degree**				
52	199	**65**	282	**2**	229	**9**	228	Frag	70.0652	C4H8N	***Small nitrogen containing fragment ion – often proline or ornithine derived – most abundant fragment in all four beers***
37	127	**182**	142	**260**	123	**193**	142	Loss	18.0080	H2O	**Water loss - indicative of a free hydroxyl group) – often seen in sugary structures**
230	99	**208**	100	**262**	90	**273**	80	Loss	46.0053	CH2O2	**Combined loss of H2O and CO – indicative for free carboxylic acid group (COOH) – generic substructure in amino acids and organic acids**
148	67	**20**	106	**195**	98	**77**	49	Frag, Frag	72.0807, 55.0546	C4H10N, C4H7	*Aliphatic amine (NH3 loss indicates free NH2 group coupled to aliphatic chain)*
24	24	**168**	25	**226**	29	**189**	25	Loss	162.0531	C6H10O5	**Loss of [hexose-H2O] – indication of hexose conjugation (for example**

											glucose)
55	14	**-**	-	**-**	-	**-**	-	Loss	74.0002	C2H3O2	*Free CO2 + CO loss, loss part of CH2O2-loss M2M for other beers*
217	61	**89**	101	**158**	66	**169**	38	Frag, Frag	86.0965, 132.1016	C5H12N, C6H14NO2	**Leucine related substructure (mzCloud) – prevalent in Beer 2**
45	39	**268**	6	**243,** **127**	30, 14	**149**	31	Frag	98.9839	H4O4P	**Fragment ion indicative for conjugation of a phosphate group (H4O4P)**
74	36	**-**	-	**-**	-	**210**	98	Frag, Frag	85.0283, 57.0332	C4H5O2, C3H5O	Two small fragments with CO loss in between. Unclear if it points to a specific substructure.
238	31	**46** **297**	20 21	**53**	25	**250**	27	Loss, Loss	179.0791, 197.0899	C6H13NO5, C6H15NO6	***Losses indicative of a hexose with NH2 group – EF fits***
129	18	**37**	22	**98**	23	**28**	22	Frag, Frag	98.0600, 144.0658	C5H8NO, C6H10NO3	*Fragment ions possibly indicative for N-Methyl-oxo-pyrrolidinecarboxylic acid like structure (loss of free carboxyl group)*
111, **270**	123, 27	**63**	147	**174,** **59**	114, 20	**170**	114	Frag, Frag, Frag	84.0442, 56.0498, 130.0505	C4H6NO, C3H6N, C5H8NO3	**Fragment ions indicative for pyroglutamic acid (pyroglutamate) or lysine (MzCloud) – structure can be formed from glutamic acid (glutamate) in the mass spectrometer as well.**
103	41	**61**	64	**214**	57	**205**	41	Loss	17.0247	NH3	**Amine loss - Indicative for free NH2 group in**

											fragmented molecule
38	39	**45**	42	**60**	40	**89**	44	Loss	36.0183	H4O2	**Double water loss, i.e., 2*H20 – Generic feature for metabolites containing several free OH groups attached to a aliphatic chain, like sugars.**
263	36	**90**	31	**151**	35	**202**	19	Frag, Loss	116.0712, 115.0630	C5H10NO2, C5H9NO2	***Fragment and loss of [proline-H2O] - indicative for conjugated proline – EF fits***
157	27	**136**	30	**280**	30	**36**	29	Loss	60.0210	C2H4O2	Loss possibly indicative of carboxylic acid group with 1-carbon attached.
264	25	**81**	12	**40**	15	**-**	-	Frag, Frag, Frag	83.0604, 56.0498, 129.0658	C4H7N2, C3H6N, C5H9N2O2	**Imidazole group linked to a carboxylgroup through one CH2 group, i.e., like in imidazole acetic acid - Prevalent in Beer1**
160	14	**298**	44	**284**	30	**-**	-	Frag, Frag, Frag	109.0288, 81.0333, 53.0888	C6H5O2, C5H5O, C4H5	**Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) – C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.**
226	75	**72**	38	**276**	20	**68**	39	Frag, Frag, Frag, Frag	105.0702, 79.0541, 91.0541, 53.0388,	C8H9, C6H7, C7H7, C4H5	**Alkyl aromatic substructure – indicative for aromatic ring with 2-carbon alkyl chain attached, i.e., phenylethene fragment from ethylbenzene as a result of the fragmentation process.**

165	56	**53**	77	**45**	56	**79**	48	Frag, Frag, Frag	84.0808, 56.0498, 67.0546	C5H10N, C3H6N, C5H7	**Fragment ions indicative for pipecolic acid (pipecolate) (MzCloud) - Quite prevalent, especially in Beer2**
131	46	**108**	41	**79,** **184**	20, 22	**243**	43	Frag, Frag, Frag	58.0655, 118.0861, 59.0733	C3H8N, C5H12NO2, C3H9N	**Fragment ions indicative for trimethylated amine connected to a carboxylic acid group, i.e., like in betaine (MzCloud)**
142	44	**6**	19	**130**	10	**69**	17	Frag, Frag, Loss, Loss	112.0511, 95.0239, 132.0419, 149.0685	C4H6N3O, C4H3N2O, C5H8O4, C5H11NO4	**Fragment ions indicative for cytosine, and a loss of conjugated deoxyribose – possibly combined due to many spectra that combine these two substructures. Loss of NH2 group is likely from remaining fragment after loss of deoxyribose. –** **Quite prevalent in Beer1**
36 **36** **36** **36**	36	**102** **102** **240** **240**	41 59	**209** **209**	48	**30** 86 + 69!)	25	Frag, Frag, Frag, Frag	114.05601, 68.0498, 69.0337, 53.0026	C4H8NO, C4H6N, C4H5O, C3HO	Possibly suggests 2-pyyrolidine substructure – Mass2Motif not consistent over the four beers.
40	26	**266**	24	**154**	26	**118**	5	Frag, Frag	71.0687, 117.0740	n/a, n/a	*C13 isotope peaks of proline (abundant ions taken for fragmentation)*
56	19	**262**	20	**34**	14	**-**	-	Loss,	88.0159,	C3H4O3,	Combination of small losses – free carboxylgroup and acetyl group loss + loss of NH2

								Loss, Loss	42.0107, 105.0425	C2H2O, C3H7NO3	group in some cases
293	18	**33**	15	**-**	-	**262**	14	Loss, Loss, Frag	60.0322, 59.0483, 60.0559	CH4N2O, CH5N3, CH6N3	Fragment and losses possibly indicative for guanidino group (CH6N3)
225	17	**51**	20	**168**	18	**258**	18	Frag, Frag, Frag	181.0970, 209.0925, 125.0708	C9H13N2O2, C10H13N2O3, C6H9N2O	Unclear yet what these fragments relate to.
33	16	**118**	11	**149**	16	**172**	15	Frag, Frag	96.0441, 124.0398	C5H6NO, C6H6NO2	Possibly suggests 2-pyridone/ol substructure
185	15	**166**	23	**220**	32	**24**	19	Frag, Frag	136.0629, 119.0351	C5H6N5, C5H3N4	**Fragments indicative adenine (C5H6N5) substructure – most prevalent in Beer3**
110	42	**116**	38	**97**	31	**29**	30	Loss, Loss, Frag, Frag, Frag	180.0632, 198.0738, 85.0283, 69.0337, 81.0334	C6H12O6, C6H14O7, C4H5O2, C4H5O, C5H5O	Oxygen-rich losses and fragments also occurring in hexose spectra – related to M2M 211 (hexose [glucose] conjugatation) – possibly hydrated-hexose loss?
279	23	**202**	22	**55**	23	**157**	27	Frag, Frag, Frag, Frag	91.0541, 119.0488, 147.0437, 65.0388	C7H7, C8H7O, C9H7O2, C5H5	**Fragments indicative for cinnamic acid (cinnamate) substructure (MzCloud)**

298	21	**77**	10	**90**	12	**98**	24	Loss, Loss, Loss, Loss	78.0316, 120.0420, 108.0421, 66.0320	C2H6O3, C4H8O4, C3H8O4, CH6O3	Combinations of small generic losses like CH2O2 + CH4O = C2H6O3
294	19	**289**	27	**241**	21	**47**	25	Frag, Frag, Frag, Frag	110.0718, 156.0769, 93.0450, 95.0608	C5H8N3, C6H10N3O2, C5H5N2, C5H7N2	**Fragments indicative for histidine (C6H10N3O2) substructure (MzCloud)**
114	18	**-**	-	**-**	-	**181**	17	Loss, Loss, Loss, Loss	59.0370, 89.0476, 42.0107, 87.0320	C2H5NO, C3H7NO2, CH2O, C4H9NO	Combinations of small generic losses like C2H2O + NH3 = C2H5NO
80	16	**-**	-	**13**	10	**-**	-	Frag, Frag	129.0658, 147.0759	C5H9N2O2, C5H11N2O3	**Fragment ions indicative for glutamine (C5H11N2O3) substructure**
177	21	**117**	50	**115**	28	**110**	26	Frag, Frag, Frag	120.0808, 103.0546, 91.0541	C8H10N, C8H7, C7H7	**Fragments indicative for [phenylalanine-CHOOH] based substructure**
67	14	**269**	18	**162**	12	**11**	13	Frag, Frag, Frag, Frag, Frag	152.0560, 153.0407, 110.0346, 135.0300, 55.0295	C5H6N5O, C5H5N4O2, C4H4N3O, C5H3N4O, C2H3N2	**Fragment ions indicative for guanine (C5H5N5O) based substructure**

195	12	**-**	-	**104**	5	**-**	-	Frag, Frag, Frag, Frag, Frag	80.0495, 164.0346, 136.0397, 53.0389, 65.0388	C5H6N, C8H6NO3, C7H6NO2, C4H5, C5H5	Unclear what these fragments relate to.
181	11	**2**	11	**19**	11	**256**	10	Frag, Frag, Frag, Frag, Frag	177.0547, 145.0284, 89.0386, 117.0331, 149.0599	C10H9O3, C9H5O2, C9H7, C8H5O, C9H9O2	**Fragments indicative for ferulic acid based substructure (MzCloud)**
22	38	**133**	50	**58**	42	**185**	55	Frag, Frag, Frag, Frag, Frag	121.0649, 103.0545, 91.0541, 53.0389, 93.0698	C8H9O, C8H7, C7H7, C4H5, C7H9	**Fragments indicative for ethylphenol substructure (i.e. resulting from Tyramine – MzCloud)**
85	37	**-**	-	**69**	47	**164**	39	Frag, Frag, Frag	69.0337, 57.0337, 73.0285	C4H5O, C3H5O, C3H5O2	***Fragment ions possibly indicative for ribose substructure (MzCloud)***
26	31	**15**	36	**7**	25	**37**	25	Frag, Frag	104.1070, 60.0810	C5H14NO, C3H10N	Possibly suggests 5-aminopentanol substructure
143	11	**-**	-	**72**	5	**-**	-	Frag,	150.0557,	C8H8NO2,	Possibly suggests methoxy-1H-indole-2,3-dione

								Frag, Frag, Frag	178.0501, 95.0494, 135.0310	C9H8NO3, C6H7O, C7H5NO2	(methoxy-isatin) substructure
245	12	**71**	20	**202**	15	**104**	9	Frag, Frag, Frag, Frag, Frag	118.0654, 117.0571, 91.0541, 130.0645 188.0706,	C8H8N, C8H7N, C7H7, C9H8N, C11H10NO2	**Fragments indicative of [tryptophan-NH3] related substructure (C8H8N is the basic indole skeleton, a fused benzene and 5 membered N-containing ring)**
244	16	**291**	22	**-**	-	**-**	-	Loss, Loss, Loss, Loss, Loss, Loss, Frag, Frag,	36.0183, 162.0525 138.0526, 196.0583, 150.0526, 64.0150, 184.0578, 112.0398, 87.0316	H4O2, C6H10O5, C7H8NO2, C6H12O7, C5H10O5, CH4O3, C5H12O7, C5H6NO2, C3H5NO2	Possibly suggests iminosugar like substructure. Losses related to sugar (polyhydroxylated structure)
146	27	**238**	9	**82**	28	**286**	8	Frag, Frag, Frag, Frag,	131.1292, 72.0809, 114.1028, 98.0600,	C5H15N4, C4H10N, C5H12N3, C5H8NO,	Possibly suggests agmatine based substructure (C5H15N4), with unknown conjugation….

								Frag, Frag, Frag, Frag	60.0559, 157.1084, 278.0554, 207.0796	CH6N3, C6H13N4O, C17H14O4, C15H11O	
5	12	**126**	13	**68**	10	**101**	7	Frag, Frag, Frag, Frag	258.1335, 276.1435, 230.1398, 212.1277	C12H20NO5, C12H22NO6, C11H20NO4, C11H18NO3	Possibly suggests iminosugar like substructure. Fragments have losses (H2O, CO) related to sugar (polyhydroxylated structure)
211	81	**111**	124	**131,** **129**	129 73 (huge overlap)	**52**	58	Frag, Frag, Frag, Frag, Frag, Frag	85.0283, 145.0550, 127.0387, 97.0284, 69.0337, 163.0605	C4H5O2, C6H9O4, C6H7O3, C5H5O2, C4H4O, C6H11O5	**Fragments indicative of a [hexose-H2O] substructure – i.e., indicative for a hexose (like glucose) conjugation (MzCloud)**
2	7	**113**	57	**102**	67	**233**	46	Frag, Frag, Frag, Frag, Frag, Frag, Frag, Frag,	67.0545, 81.0700, 55.0540, 149.1325, 277.2173, 295.2288, 93.0698, 71.0857,	C5H7, C6H9, C4H7, C11H17 C18H29O2, C18H31O3, C7H9, C5H11,	Possibly suggests alkylbenzene substructure.

								Frag, Frag	141.1273, 169.1226	C9H17O, C10H17O2	
166	9	**227**	19	**121**	10	**49**	8	Frag, Frag, Frag, Frag	146.0811, 128.0703, 81.0334, 83.0490	C6H12NO3, C6H10NO2, C5H5O, C3H7O	Possibly suggests 4-aminooxane-4-carboxylic acid like substructure?
-	-	**75**	14	**-**	-	**46**	12	Frag, Frag, Frag	86.0314, 146.0528, 128.0428	N/A N/A N/A	Isotope M2M of 111 (glycoside/hexoside related)
-	-	**162**	86	**176**	57	**129**	80	Frag	91.0541	C7H7	*Small abundant and generic aromatic fragment found across several mass patterns.*
-	-	**240**	59	**290**	69	**191**	72	Frag, Frag	69.0701, 53.0026	C5H9, C3HO	Two small fragments, unclear if they represent a substructure
-	-	**217**	29	**-**	-	**-**	-	Loss	35.0343 (35.0366!)	H5NO	***Combined (sequential) H2O and NH3 loss***
184	9	**36**	13	**207**	13	**116**	17	Frag, Frag	152.0703, 134.0600	C8H10NO2, C8H8NO	Unclear what these fragments relate to.
-	-	**88**	62	**221**	42	**263,** **0**	68, 13	Frag, Frag, Frag, Frag	57.0701, 85.0648, 67.0546 53.0026	C4H9, C5H9O, C5H7, C3HO	Unclear yet what these fragments relate to.
-	-	**260**	29	**233**	16	**137**	18	Loss,	64.0161,	CH4O3,	Combination of small losses

								Loss	92.0108	C2H4O4	(CO2, H2O, etc.) – Unclear if they relate to a substructure loss.
-	**-**	**134**	18	**222**	11	**-**	-	Frag, Frag, Frag, Loss	60.0448, 106.0497, 88.0392 115.0268	C2H6NO, C3H8NO3, C3H6NO2, C4H5NO3	**Fragments (and loss) indicative for serine substructure (MzCloud) -** **Present in Beer 2 & Beer 3**
-	**-**	**243**	16	**-**	-	**-**	-	Loss, Loss	143.0580, 99.0682	C6H9NO3, C5H9NO	Unclear what these fragments relate to.
-	**-**	**187**	30	**230**	31	**-**	-	Frag, Loss, Frag, Frag	87.0439, 86.0366, 104.0711, 69.0337	C4H7O2, C4H6O2, C4H10NO2, C4H5O	**Fragments indicative for y-aminobutyric acid (amimobutyrate) substructure (MzCloud) – present in Beer 2 & Beer 3 – in beer 3 mainly based on C4H7O2 fragment.**
93	34	**259**	41	**12**	44	**229**	32	Loss, Loss, Loss, Loss	27.9941, 30.0100 55.9897, 54.0102	CO, CH2O C2O2, C3H2O	Combination of small losses (CO2, H2O, etc.) – Unclear if they relate to a substructure loss.
-	**-**	**10**	32	**227**	29	**151**	53	Frag, Frag, Frag	111.0443, 83.0490, 55.0547	C6H7O2, C5H7O, C4H7	Possibly related to 1,4-Cyclohex-2-enedione substructure – double CO loss between fragments. Could be alkaloid fragments as well.
62	13	**66**	24	**136**	19	**66**	15	Frag, Frag	95.0607, 68.0498	C5H7N2, C4H6N	Unclear what these fragments relate to. Possibly small ring

											structure (CHN loss)
107	9	**263**	16	**-**	-	**-**	-	Frag, Loss, Loss, Frag	128.1074, 60.0576, 42.0470, 110.0970	C7H14NO C2H4O2, C2H2O, C7H12N	Unclear what these fragments relate to.
6	6	**290**	11	**67**	3	**67**	5	Frag, Frag, Frag, Frag	68.9972, 111.0076, 129.0186, 157.0131	C3HO2, C5H3O3, C5H5O4, C6H5O5	**Fragment ions indicative for aconitic acid substructure (C3HO2 fragment is quite specific)**
-	-	**224**	10	**-**	-	**71**	10	Loss, Loss	53.0476, 71.0583	NH7O2, NH9O3	Combination of small losses (NH3, H2O) – Unclear if they relate to a substructure loss.
-	-	**153**	19	**294**	11	**-**	-	Frag, Frag	180.1013, 162.0915	C10H14NO3, C10H12NO2	Unclear what these fragments relate to.
281	9	**77**	10	**249**	12	**26**	11	Frag, Frag, Frag, Frag	138.0545, 140.1065, 186.0758, 168.0650	C7H8NO2, C7H10NO2, C8H12NO4, C8H10NO3	Unclear yet what these fragments relate to.
125	13	**246**	18	**17**	6	**8**	10	Frag, Frag, Frag, Frag, Frag,	136.0760, 107.0493, 91.0543, 95.0494, 123.0447,	C8H10NO, C7H7O, C7H7, C6H7O, C7H7O2,	**Fragments indicative for tyrosine substructure (MzCloud)**

								Frag, Frag	119.0488, 182.0822	C8H7O, C9H12NO3	
228	5	**200**	10	**4**	6	**121**	11	Frag, Frag, Frag, Frag	260.1117, 128.0704, 242.1011, 100.0754	C11H18NO6, C6H10NO2, C11H16NO5, C5H10NO	Unclear yet what these fragments relate to.
-	-	**173**	25	**41**	32	**128**	5	Frag, Frag, Frag, Frag, Frag	130.0506, 97.0284, 238.0714, 226.0718, 274.0920	C5H8NO3, C5H5O2, C11H12NO5, C10H12NO5, C11H16NO7	Unclear yet what these fragments relate to.
-	-	**254**	15	**188**	10	**42**	14	Frag, Frag, Frag, Frag	73.0285, 133.0499, 57.0337, 115.0391	C3H5O2, C5H9O4, C3H5O, C5H7O3	Unclear yet – possibly related to methylsuccinic acid….
-	-	**-**	-	**128**	15	**-**	-	Loss	42.0107	C2H2O	N/O-Acetylation (Beer 3)
197	14	**67**	22	**250**	22	**-**	-	Loss, Loss, Loss	63.0319, 45.0578, 91.0268	CH5NO2, C2H7N, C2H5NO3	Combination of small losses (i.e., NH3 and CH2O2)
-	-	**96**	18	**272**	12	**199**	14	Frag, Frag	74.0598, 56.0497	C3H8NO, C3H6N	Unclear if fragments relate to a specific substructure.
-	-	**78**	14	**291**	22	**173**	12	Frag,	55.0547,	C4H7,	Unclear if fragments relate to a specific substructure. H2O

								Frag	73.0647	C4H9O	loss between fragments.
-	-	**-**	-	**139,** **180**	20, 10	**132**	15	Frag, Frag, Frag, Frag	89.0600, 133.0863, 177.1128, 111.0443	C4H9O2, C6H13O3, C8H17O4, C6H7O2	Unclear yet what these fragments relate to.
286	6	**145**	9	**42**	17	**-**	-	Frag, Frag, Frag, Frag, Frag, Frag	74.0235, 88.0392, 70.0290, 87.0554, 133.0615, 116.0344	C2H4NO2, C3H6NO2, C3H4NO, C3H7N2O, C4H9N2O3, C4H6NO3	**Fragments indicative for asparagine substructure (MzCloud) – prevalent in Beer 3**
7	10	**165**	24	**91**	14	**84**	34	Frag, Frag, Frag	108.0443, 80.0495, 53.0389	C6H6NO, C5H6N, C4H5	Fragments possibly suggest benzene ring substituted with one hydroxyl and one NH2 group (fragments point to orientation from 3-hydroxyanthranilic acid – i.e. MzCloud) – prevalent in Beer 2 and 4
124	7	**4**	19	**166**	10	**78**	6	Frag, Frag, Frag	126.0665, 109.03976, 108.0560	C5H8N3O, C5H5N2O, C5H6N3	**Fragment ions indicative for 5-methylcytosine substructure (MzCloud) – prevalent in Beer 2**
130	11	**29**	18	**211**	24	**181**	17	Loss, Frag, Frag,	59.0370, 114.0912, 72.0447,	C2H5NO, C6H12NO, C3H6NO,	**Fragment ions indicative for N-acetylputrescine substructure (MzCloud)**

								Frag	60.0448	C2H6NO	
-	**-**	**92**	27	**156**	22	**99**	36	Loss	132.0421	C5H8O4	***[Ribose (pentose, C5-sugar)-H2O] related loss – indicative for conjugated ribose sugar - EF fits***
201	9	**185**	8	**270**	8	**246**	12	Frag, Frag, Frag	206.1024, 86.0602, 74.0600	C8H16NO5, C4H8NO, C3H8NO	Unclear yet what these fragments relate to.
77	10	**135**	16	**23**	16	**244**	18	Loss, Loss, Loss	144.04192, 190.0474, 160.0370	C6H8O4, C7H10O6, C6H8O5	Unclear yet what these losses relate to.
-	**-**	**13**	10	**-**	-	**282**	14	Frag, Frag	126.0600, 94.0648	C7H8NO, C6H8N	Unclear yet what these fragments relate to.
-	**-**	**-**	-	**35**	11	**161**	18	Frag, Frag	95.0494, 137.0600	C6H7O, C8H9O2	Unclear yet what these fragments relate to.

Table S-5. Table with Mass2Motifs (MSMs) discover in the four negative ionization mode fragmentation files of the beer extracts.

Table with M2Ms in four beers – fragments/losses associated to the M2Ms can slightly differ in between beers due to degree and type of metabolites associated to the M2M. Experimental masses are within 5 ppm of theoretical masses as found in topics. Slight changes are observed per file. Annotated Mass2Motifs: Bold represents highest level of confidence (i.e., several fragments or specific mass value that can only point to a certain combination of ions), bold and italic is second-highest level of confidence (i.e., match on elemental formula (EF) only – but in the given sample matrix it is quite a likely structural annotation), just italic is the third-highest level of confidence (i.e., no specific structure found, often generic fragments that have multiple possible structural confirmations), and plain text represents the lowest level of confidence.

Beer1		**Beer2**		**Beer3**		**Beer4**					
								Frag/ Loss	**m/z**	**EF**	**Characterization**
M2M	**Degree**	**M2M**	**Degree**	**M2M**	**Degree**	**M2M**	**Degree**				
0	161	**198**	108	**74**	156	**84**	126	Frag	71.0135	C3H3O2	***Fragment ion related to 3-hydroxy-carboxilic acid substructure (C=C=O coupled to C-O[-]) - EF fits***
147	83	**133**	31	**158**	49	**104**	41	Frag	101.0248	C4H5O3	***2-oxo-butyric acid (2-oxo-butyrate) fragment - EF fits***
86	84	**205**	90	**75**	68	**25**	73	Loss	43.9898	CO2	**Loss of carboxilic acid group - suggests free CO2 group (for example in underivatized amino acid)**
287	66	**48**	68	**273**	67	**281**	44	Frag	85.0295	C4H5O2	*Fragment related to small organic acid - usually contains carboxylic acid group with (branched/unbranched) 3-carbon alkylchain attached to it.*
116	66	**240**	81	**50**	71	**253**	49	Frag	78.9593	PO3	**Fragment of phosphonate - indicates phosphate**

											substructure
233	56	**284**	59	**180**	49	**273**	32	Frag	59.0133	C2H3O2	***Fragment consisting of aldehyde and hydroxyl group - common structural motif in sugar fragmentation - EF fits***
54	54	**-**	-	**-**	-	**-**	-	Frag	80.9649	HSO3	**Fragment of sulphate anion, fragmented from aliphatic chain -** **Only present in Beer 1**
137	137	**184**	43	**292**	34	**105**	30	Loss	162.0529	C6H10O5	**Loss of [hexose-H2O] - indication of hexose conjugation (for example glucose)**
257	40	**157**	48	**82**	48	**9**	39	Loss	18.0094	H2O	**Loss of water molecule (H2O) - indication of free hydroxyl group**
5	36	**50**	40	**47**	23	**74**	33	Loss	62.0005	CH2O3	*Combined loss of CO2 and H20, possibly suggests two carboxylic acid groups in the fragmented metabolite*
156	25	**30**	18	**201**	26	**133**	14	Loss	72.0212	C3H4O2	Loss possibly indicative of carboxylic acid group with 2-carbon alkyl chain attached.
230	23	**259**	17	**246**	20	**196**	17	Loss	60.0210	C2H4O2	Loss possibly indicative of carboxylic acid group with 1 carbon attached.
163	23	**28**	24	**145**	12	**43**	28	Frag	87.0086	C3H3O3	***Fragment related to pyruvic acid (pyruvate) or oxaloacetate - EF fits***

111	22	**263**	21	**162**	22	**60**	18	Loss	90.0318	C3H6O3	***Loss related to lactic acid (lactate) - EF fits***
65	20	**108**	14	**256**	11	**158**	12	Loss	71.9849	C2O3	*CO2 loss and CO loss combined – not clear if this points to a substructure*
63	19	**110**	19	**13**	16	**121**	13	Loss	116.0111	C4H4O4	*Loss possibly indicative of fumaric acid (fumarate) - EF fits*
72	20	**297**	15	**88**	24	**69**	29	Frag	60.9927	CHO3	*Bicarbonate fragment - possibly related to small oxygen rich organic acids*
195	15	**-**	-	**113**	9	**-**	-	Frag	69.0343	C4H5O	*Fragment ion indicative for carboxylic acid group with a 3-carbon alkyl chain attached.*
286	13	**83**	21	**184**	11	**183**	14	Frag, Loss	114.0558, 115.0637	C5H8NO2, C5H9NO2	***Fragment and loss related to proline substructure - EF fits***
38	6	**79**	18	**54**	20	**72**	19	Loss	46.0057	CH2O2	Combined losses of H2O and CO - not sure if this relates to a particular structural feature
6	67	**131**	8	**216**	11	**264**	8	Frag, Frag	79.9575	SO3	**Fragment of sulphate ion, fragmented from aromatic structure**
167	53	**228**	54	**101**	45	**148 and 221**	29 11	Frag	89.0249, 71.0136	C3H5O3, C3H3O2	**Fragments indicating lactic acid (lactate) substructure (MzCloud)**
141	43	**85**	44	**56**	40	**151**	27	Frag	128.0358	C5H6NO3	Generic fragment - unclear if any specific substructure is related

178	36	**226**	42	**105**	35	**118**	33	Frag	88.0407	C3H6NO2	***Fragment ion indicating alanine substructure - EF fits***
139	34	**21**	48	**284**	30	**67**	23	Frag, Frag	94.0301, 66.0346	C5H4NO, C4H4N	***Fragments related to nicotinic acid (nicotinate) substructure - MzCloud***
152	32	**298**	24	**32**	37	**296**	37	Frag	72.9928	C2HO3	**Fragment related to 2-hydroxycarboxilic acid related substructure - indicative for a carboxylic acid group with one carbon attached bearing a hydroxyl group**
297	23	**71**	24	**254**	28	**194**	20	Frag	75.0085	C2H3O3	***Fragment related to 2-hydroxyethanoic acid substructure - MzCloud***
255	21	**98**	32	**164**	26	**92**	14	Loss	180.0655	C6H12O6	*Loss possibly indicating hydrated hexose loss*
217	19	**3**	10	**168**	17	**298**	19	Loss	27.9945	CO	**Loss of C=O - small loss, unclear what it points to in negative ionization mode**
214	11	**277**	13	**130**	21	3	5	Loss	129.0428	C5H7NO3	Loss possibly related to pyroglutamic acid (pyroglutamate) - EF fits
42	33	**-**	-	**-**	-	-	-	Frag, Frag	179.0572, 161.0465	C6H11O6, C6H9O5	***Fragments suggesting hexose substructure - EF fits – in Beer1 only***
271	28	**215**	24	**62**	28	**101**	14	Frag	74.0245	C2H4NO2	***Glycine related fragment - EF fits***
284	26	**213**	18	**174**	22	**131**	16	Frag	73.0294	C3H5O2	***Fragment indicative for ethylcarboxylate substructure***

											- MzCloud
128	27	**156**	56	**225**	23	**98**	16	Frag	130.0881	C6H12NO2	***Fragment indicative of leucine substructure - EF fits - MzCloud***
298	15	**38**	11	**21**	13	**227**	13	Loss	87.9797	C2O4	**Loss of two CO2 molecules – indicative for two free carboxylic acid groups**
254	16	**54**	11	**288**	8	**252**	13	Loss	132.0423	C5H8O4	***Loss indicating [pentose (C5-sugar)-H2O] loss - indicative for conjugated pentose sugar - EF fits***
36	12	**199**	14	**161**	18	**146**	12	Frag	102.0564	C4H8NO2	Fragment possibly suggesting aminobutyric acid (aminobutyrate) substructure
119	33	**-**	**-**	**-**	**-**	**214**	14	Frag, Frag, Frag, Frag	72.9928, 59.0134, 119.0348, 91.0404	C2HO3, C2H3O2, C4H7O4, C3H7O3	*Fragments possibly related to threose substructure*
76	16	**174**	17	**237**	12	**162**	12	Frag	127.0510	C5H7N2O2	***Fragment indicative of glutamine substructure - EF fits with [glutamine-COOH]***
129	15	**195**	17	**187**	12	**73**	4	Loss, Loss	120.0423, 108.0425	C4H8O4, C3H8O4	Losses possibly related to small sugar like threose
1	7	**229**	6	**120**	16	**123**	5	Frag, Frag, Frag,	545.1700, 383.1197, 221.0656,	C8H13O7,	***Fragments related to polysaccharides - this mass2motif contains doubly charged species - it is unclear whether that points to a specific structural feature of***

								Frag, Frag	1031.3366, 161.0448	 C6H9O5	***the polysaccharide structure***
69	17	**90**	14	**136**	17	**192**	19	Frag, Frag, Frag, Frag	383.1197, 161.0439, 545.1700, 221.0684	 C6H9O5 C8H13O7	***Fragments related to polysaccharides - this mass2motif contains singly charged species - it is unclear whether that points to a specific structural feature of the polysaccharide structure***
1	7	**14**	14	**291**	11	**-**	-	Frag, Frag, Frag, Frag, Frag	221.0684, 179.0571, 161.0464, 119.0348, 85.0294	C8H13O7, C6H11O6, C6H9O5, C4H7O4, C5H5O2	***Fragments related to polysaccharides - this mass2motif contains just the smaller m/z fragments with C8H13O7 as largest fragment, indicative for a disaccharide***
226	9	**115**	13	**95**	6	**235**	7	Frag, Frag, Frag, Frag, Frag, Frag	150.0420, 133.0157, 126.0316, 151.0472 66.0097, 108.0209	C5H4N5O, C5HN4O, C4H4N3O2, C2H7N4O4 C2N3, C4H2N3O	**Fragments indicative for guanine (C5H4N5O) substructure – (MzCloud)**
25	7	**275**	28	**199**	12	**33** (no 93 Fragment)	9	Frag, Frag, Frag, Loss,	93.0349, 191.0560, 173.0456, 174.0532,	C6H5O, C7H11O6, C7H9O5, C7H10O5,	**Fragments indicative for caffeoylquinic acid like metabolites - prevalent in Beer 2**

								Frag	137.0616	C7H5O3	
202	7	**230**	8	**-**	-	**-**	-	Frag	125.0365	C5H5N2O2	Fragment possibly suggests imidazoleacetic acid substructure – EF fits
-	-	**180**	15	**191**	20	**-**	-	Frag, Frag	97.0296, 69.0343	C5H5O2, C4H5O	***Fragments indicative for polyhydroxylated benzene ring (e.g. pyrogallol)***
-	-	**239**	19	**128**	35	**18**	23	Frag	161.0464	C6H9O5	Fragment related to hexose - unclear if it points to specific structural feature
-	-	**258**	9	**271**	18	**-**	-	Frag	179.0572	C6H11O6	Fragment related to hexose - unclear if it points to specific structural feature
9	12	**144**	12	**287**	12	**245**	5	Frag, Frag	111.0084, 173.0090	C5H3O3, C6H6O6	**Fragments indicative for citric acid (citrate) substructure - (MzCloud)**
132	16	**123**	31	**2**	14	**41**	46	Frag	125.0605	C7H9O2	Fragment unclear yet what this points to - predominant in Beer 2 and 4.
-	-	**288**	17	**106**	16	**229**	21	Frag	111.0451	C6H7O2	*Fragment possibly indicative for carboxylic acid group with a 5-carbon alkyl chain attached.*
58	14	**181**	16	**193**	14	**271**	14	Loss	42.0103	C2H2O	**Loss of acetyl group - indicative for the conjugation of acetic acid.**
220	12	**148**	10	**226**	13	**202**	25	Frag, Frag	59.9849, 56.9952	CO3, ??!!	Fragment possibly representing bicarbonate anion - unclear if this points to a

											structural feature
3	15	**219**	13	**87** And **99**	5 8	**66**	10	Frag, Frag	164.0716, 147.0452	C9H11NO2, C9H8O2	**Fragments indicative to phenylalanine substructure**
34	11	**65**	13	**140**	16	**38**	18	Frag	119.0508	C8H7O	*Fragment possibly suggests hydroxyphenylethylene substructure*
-	-	**-**	-	**133**	14	**297**	12	Frag, Frag, Frag	124.0400, 94.0301, 66.0346	C6H6NO2, C5H4NO, C4H4N	Fragments unclear yet to which substructure they relate - related to nicotinate substructure fragments
7	9	**3**	7	**2**	12	**86**	9	Frag, Frag	96.9599, 79.9575	HSO4, SO3	***Fragments indicative for sulphate group substructure - unclear if there is a specific configuration that results in the HSO4 fragment***

S2.2 Feature Extraction in the MS2LDA Workflow

All fragmentation files from the four Beers, including fragmentation files of the pooled beer sample, were run through the Data Conversion part of MS2LDA. Here, we explored three alternative methods for linking MS2 spectra to MS1 peaks that were picked by XCMS after the Peak Detection step in the MS2LDA workflow. These three methods, labelled by their numbers in the list below, can be described as follows:

1. This method uses an XCMS function (xcmsFragments) on the same fragmentation file for both MS1 peak picking and finding correspondent MS2 spectra.
2. This method is based on a modified xcmsFragments script that uses both a full scan file for MS1 peak picking and a separate fragmentation file for finding correspondent MS2 spectra.
3. This method is similar to method 2 in that it uses two separate full-scan and fragmentation files for the MS1 peak picking and finding correspondent MS2 spectra, but it is based on the RMassBank scripts for MS1-MS2 pairing (2).

Method 3 was also tested using different sources of fragmentation spectra, namely from the pooled beer sample run with the combined fragmentation mode, and the separate fragmentation mode, as well as from the corresponding sample, in both fragmentation modes.

The following Table S-6 shows the number of mass features extracted by XCMS and number of unique MS1-MS2 pairs (picked MS1 peaks that were fragmented at least once during the fragmentation run) found for the eight files used in the study and for the different MS1-M2 pairing methods. Table S-3 also shows that using method 3 and the fragmented sample (in ‘Separate Fragmentation Mode’, i.e., using one ionization mode) as source of fragmentation spectra, half of the detected features above 3E5 cts have an MS2 spectrum matched.

	XCMS - total MS1 features above 3E5 cts	Total MS2 spectra within RT window 3-21min	Unique MS1-MS2 pairs Meth 1	Unique MS1-MS2 pairs Meth 2	Unique MS1-MS2 pairs Meth 3 Pooled Combined	Unique MS1-MS2 pairs Meth 3 Pooled Separate	Unique MS1-MS2 pairs Meth 3 Sample Combined	Unique MS1-MS2 pairs Meth 3 Sample Separate
Beer1POS	3136	5474	700	933	817	1297	878	1282
Beer2POS	3439	5499	808	1107	858	1403	818	1567
Beer3POS	3268	5457	737	999	835	1320	832	1422
Beer4POS	3222	5189	707	1004	764	1255	820	1363
Beer1NEG	1980	4540	349	459	555	752	620	1178
Beer2NEG	2082	4486	423	466	568	789	591	1178
Beer3NEG	1932	4335	394	492	532	704	532	1126
Beer4NEG	1807	4242	382	428	492	705	544	1018

Table S-6. Number of mass features extracted by XCMS and number of unique MS1-MS2 pairs (picked MS1 peaks that were fragmented at least once during the fragmentation run) found for the eight files used in the study and for the different MS1-M2 pairing methods.

S2.3 Mass2Motifs and MS1 Peaks Statistics

On average, ~70% of fragmentation spectra can be explained by at least one structurally annotated Mass2Motifs (Table S-7).

File	Total MS1 peaks fragmented	MS1 peaks linked to at least one structurally annotated M2M	%
Beer1POS	1282	951	**74**
Beer2POS	1567	1160	**74**
Beer3POS	1422	1055	**74**
Beer4POS	1363	930	**68**

Table S-7. Mass2Motif coverage of MS1 peaks by percentage of MS1 peaks that can be explained by at least one structurally annotated Mass2Motif for the files acquired in positive ionization mode.

S2.4 Metabolite Annotations Using Mass2Motif Membership and Spectral Matching to the Nist_msms and MassBank Databases

To assess how MS2LDA contributes to metabolite annotation, the MS1 peaks associated to the structurally characterized Mass2Motifs related to ferulic acid (M2M_19), histidine (241), tyrosine (17) and tryptophan (202) in the Beer3 POS file were analysed in detail. Metabolite annotations were done using the structural information provided by MS2LDA. The resulting annotations can be found in Table S-8. Please note that most of those metabolites are no peptides thus representing small molecules differently from those encountered in proteomics/peptidomics and that out of the 51 associated MS1 peaks 9 were incorrectly associated to a

particular Mass2Motif by co-elution and co-fragmentation with an isobaric species that does genuinely contain the Mass2Motif substructure. To remove such incorrect associations, further improvements to obtain clear fragmentation spectra for each metabolite would be needed. Also, a fragment and an isotope were included in the associated MS1 peaks for histidine, leaving 39 metabolite features for further analysis.

Table S-8. Metabolite annotations based on Mass2Motif membership. * indicates doubly charged species. The most likely annotation is presented based on Mass2Motif membership (classification) and the corresponding Metabolomics Standards Initiave Metabolite Identification level is indicated. The last column indicated whether or not the mass was annotated with a peptide.

M2M	**Mass $[M+H]^+$**	**EF $[M+H]^+$ (most likely)**	**RT (s)**	**Class**	**Annotation**	**MSI MI level**	**Peptide?**
19	540.3306	C30H44N4O5	276	-	Co-elution and Co-fragmentation	4	-
19	307.1767	C15H23N4O3	547	Ferulic acid	Feruloylagmatine	3	No
19	540.2707	C29H38N3O7	263	Ferulic acid	Diferuloyl-N1-acetylspermidine	3	No
19	498.2599	C27H36N3O6	616	Ferulic acid	Diferuloyl-spermidine	3	No
19	307.0998	C8H15N6O7	613	-	Co-elution and Co-fragmentation	4	-
19	369.1182	C17H21O9	296	Ferulic acid	Feruloylquinic acid	3	No
19	314.1386	C18H20NO4	270	Ferulic acid	Feruloyltyramine	3	No
19	265.1545	C14H21N2O3	1101	Ferulic acid	Feruloylputrescine	3	No
19	194.0812	C10H12NO3	364, 378	Ferulic acid	Feruloylamine	3	No
19	195.1130	C10H15N2O2	379	-	Co-elution and Co-fragmentation	4	-
241	277.1582*	C34H43N4OP	305	Histidine	[Histidine-COOH] substructure present in molecule	3	No
241	318.1295	C12H20N3O7	569	Histidine	Histidine-hexoside	3	No
241	480.1822	C18H30N3O12	600	Histidine	Histidine-dihexoside	3	No
241	277.1585*	C34H43N4OP	427	Histidine	[Histidine-COOH] substructure present in molecule	3	No
241	310.2125	C16H28N3O3	240	Histidine	Histidine-decanoate conjugate	3	No
241	156.0768	C6H10N3O2	621	Histidine	Histidine	1	No
241	198.0873	C8H12N3O3	481	Histidine	Acetyl-histidine	3	No
241	364.1614	C16H22N5O5	466	Histidine	Histidine containing	3	Possibly

					metabolite		
241	553.3097	C34H43N4OP	305	Histidine	[Histidine-COOH] substructure present in molecule – singly charged species of 277.1582 RT 305	3	No
241	362.2166	C15H30N4O6	503	Histidine	Histidine-deoxy-trimethylamino-hexoside [conjugate of C9H22NO5-H2O]	3	No
241	110.0713	C5H8N3	621	Histidine	Histidine fragm.	-	-
241	235.1077	C12H15N2O3	409	-	Fragments of histidine motif co-fragmented by co-elution	4	-
241	235.1187	C11H15N4O2	414, 398	Histidine	[Histidine-COOH] substructure present in molecule - Possibly Histidine-C5H5N conjugate	3	No
241	157.0801	C5[C13]H10N3O2	621	Histidine	Histidine isotope	-	-
241	277.1474	?	435	Histidine	Fragments from motif in MS2 spectrum	3	-
241	251.1499	C12H19N4O2	409	Histidine	[Histidine-COOH] substructure present in molecule - Possibly Histidine-C6H9N conjugate	3	No
241	195.0876	C8H11N4O2	511	Histidine	[Histidine-COOH] substructure present in molecule	3	No
241	157.0738	C7H11NO3	621	-	Fragments of histidine motif co-fragmented by co-elution	4	-
241	272.0876	C10H16N3O6	592	Histidine	[Histidine-COOH] substructure present in molecule – Conjugated with [C5H8O7-H2O]	3	No
241	363.1760	?	904	-	Only one fragment of Mass2Motif present in MS2 spectrum	4	-
17	293.1131	C14H17N2O5	431	Tyrosine	Pyroglutamyl-Tyrosine	3	Yes

17	182.0812	C9H12NO3	585	Tyrosine	Tyrosine	1	No
17	280.1543	C15H22NO4	255	Tyrosine	Tyrosine-hexanoate conjugate (or structural isomer of [C6H12O2-H2O]	3	No
17	308.1856	C17H26NO4	234	Tyrosine	Tyrosine-octanoate conjugate (or structural isomer of [C8H16O2-H2O]	3	No
17	161.0921	C6H13N2O3	385	-	Not related to Tyrosine	4	-
17	154.0974	C8H12NO2	417	-	Not related to Tyrosine – one abundant fragment in common	4	-
202	205.0972	C11H13N2O2	554	Tryptophan (indole)	Tryptophan	1	No
202	206.0811	C11H12NO3	414	Tryptophan (indole)	3-Indolelactate (analogue of Tryptophan with NH2 replaced by OH; sharing the same indole backbone)	3	No
202	367.1500	C17H23N2O7	504	Tryptophan (indole)	Tryptophan-hexoside	3	No
202	218.0811	C12H12NO3	279	Tryptophan (indole)	3-Indoleoxobutyrate	3	No
202	334.1398	C16H20N3O5	541	Tryptophan (indole)	Glutamyl-Tryptophan	3	Yes
202	188.0706	C11H10NO2	553	Tryptophan (indole)	Fragment of Tryptophan	-	-
202	291.0973	C14H15N2O5	561	Tryptophan (indole)	Indole containing molecule - cofragmentation	3	No
202	222.1124	C12H16NO3	288	Tryptophan (indole)	Indole containing molecule	3	No
202	277.1585*	C34H43N4OP	427	Tryptophan (indole)	Indole containing molecule	3	No
202	277.1474	?	436	-	Cofragmentation	4	-
202	237.0869	C11H13N2O4	470	Tryptophan	Indole containing	3	No

				(indole)	molecule		
202	208.0597	C10H12NO4	445	Tryptophan (indole)	Indole containing molecule	3	No
202	261.0934	?	508	-	Few low abundant fragments related to indole	4	-
202	190.1437 (190.0861)	C9H20NO3 (C11H12NO2)	435	Tryptophan (indole)	Co-elution with: 3-Indolepropionic acid	3	No
202	146.0599	C9H8NO	410	-	Fragment of 3-Indolepropionic acid	-	-

In order to assess how well spectral matching would perform on the same set of metabolites annotated based on their Mass2Motif (see Table S-8). Spectral matching was performed using the mspepsearch program (http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:mspepsearch) against a local instance of the Nist_msms and the MassBank databases. For every fragmentation spectra in all the Beer datasets, an .MSP file was generated. This file was used as input for spectral matching using mspepsearch against the two spectral databases. The results from spectral matching were stored and used to obtain the matches for the metabolites from Table S-8 by specifying m/z and RT tolerances for the parent (MS1) peaks to search for, or by specifying a Mass2Motif ID (number). In the latter case, spectral annotations of all fragmentation spectra that can be explained by that Mass2Motif (at above the threshold on the Mass2Motif-to-spectra distributions) can be retrieved. Table S-9 presents the results of the spectral matching of ferulic acid, histidine, tyrosine, and tryptophan related metabolites, showing that out of the 39 with MS2LDA annotated metabolites, 7 resulted in correct hits with another 8 producing structurally related hits. These results clearly demonstrate the annotative power of MS2LDA, through which annotations can be made by matching only small portions of the spectra and therefore allowing annotation (classification) of molecules not present in database.

Table S-9. Results of spectral matching of MS1-MS2 pairs explained by four Mass2Motifs. Masses for which a correct match was found in any of the three databases (Nist_msms, Nist_msms2, and MassBank) are indicated in bold and * indicates doubly charged species. Mass2Motif numbers correspond to the beer3 positive ionization data set.

M2M	**Mass $[M+H]^+$**	**EF $[M+H]^+$ (most likely)**	**RT (s)**	**Database Annotations**	**EFs top hits of each database**	**Database**	**Normalized Score**
19	540.3306	C30H44N4O5	276	Co-elution and Co-fragmentation	-	-	-
19	307.1767	C15H23N4O3	547	Pinolenic acid ethyl ester	C20H34O2	Nist_msms	6.79
19	540.2707	C29H38N3O7	263	Leptomycin B Anti-Inflammatory Peptide 1|	C33H48O6 C45H82N12O14S2	Nist_msms Nist_msms2	32.3 20.29
19	498.2599	C27H36N3O6	616	Chicoric acid (2R,3R-O-dicaffeoyltartaric	C22H18O12	Nist_msms	79.24

				acid)			
19	307.0998	C8H15N6O7	613	Co-elution and Co-fragmentation	-	-	-
19	369.1182	C17H21O9	296	Curcumin	C21H20O6	Nist_msms	67.72
19	314.1386	C18H20NO4	270	Stearic acid ethyl ester	C20H40O2	Nist_msms	19.87
19	265.1545	C14H21N2O3	1101	3,4-Dihydroxycinnamic acid (L-alanine methyl ester) amide	C13H15NO5	Nist_msms	86.02
19	194.0812	C10H12NO3	364, 378	3-Hydroxy-4-methoxycinnamic acid (=ferulic acid) Prowl(TM)	C10H10O4 C13H19N3O4	Nist_msms MassBank	87.71 3.75
19	195.1130	C10H15N2O2	379	Co-elution and Co-fragmentation	-	-	-
241	277.1582*	C34H43N4OP	305	Leu-Enkephalin, amide|1/4,L,Amidated PyroGlu-Phe	C28H38N6O6 C14H16N2O7	Nist_msms2 Nist_msms	13.21 9.33
241	318.1295	C12H20N3O7	569	5(S),6(R)-11-trans DiHETE Tyr-His	C20H32O4 C15H10N4O4	Nist_msms Nist_msms	93.28 0.71
241	480.1822	C18H30N3O12	600	1-(9Z-Octadecenoyl)-sn-glycero-3-phosphoethanolamine	C23H46NO7P	Nist_msms	48.9
241	277.1585*	C34H43N4OP	427	L-Saccharopine Leu-Enkephalin, amide|1/4,L,Amidated L-Saccharopine	C11H20N2O6 C28H38N6O6 C11H20N2O6	MassBank Nist_msms2 Nist_msms	7.79 6.28 5.79
241	310.2125	C16H28N3O3	240	Sar1,Ala8] Angiotensin II|1/0,G,N-	C43H67N13O10	Nist_msms2	81.08

				Methyl|76/76 D-erythro-Sphingosine C-20	C20H41NO2	Nist_msms	17.27
241	**156.0768**	C6H10N3O2	621	His L-Histidine	C6H9N3O2 C6H9N3O2	MassBank Nist_msms	98.35 98.35
241	**198.0873**	C8H12N3O3	481	N-Acetylhistidine His-Leu-Lys	C8H11N3O3 C18h32N6O4	MassBank Nist_msms	98.94 0.98
241	364.1614	C16H22N5O5	466	pyro-Glu-His-Pro-NH2 TRH (Protirelin)	C16H22N6O4 C16H22N6O4	Nist_msms MassBank	89.01 89.01
241	553.3097	C34H43N4OP	305	Inosine 5'-triphosphate R15K, HIV-1 Inhibitory Peptide||26/26	C10H15N4O14P3 C73H126N26O18	Nist_msms Nist_msms2	21.35 17.2
241	362.2166	C15H30N4O6	503	pyro-Glu-His-Pro-NH2 TRH (Protirelin)	C16H22N6O4 C16H22N6O4	Nist_msms MassBank	89.14 89.14
241	110.0713	C5H8N3	621	- (fragment)	-	-	-
241	235.1077	C12H15N2O3	409	His-Pro	C11H16N4O3	Nist_msms	98.98
241	235.1187	C11H15N4O2	414, 398	His-Pro His-Pro	C11H16N4O3 C11H16N4O3	Nist_msms Nist_msms	98.98 98.98
241	157.0801	C5[C13]H10N3O2	621	- (isotope)	-	-	-
241	277.1474	?	435	L-Saccharopine Leu-Enkephalin, amide|1/4,L,Amidated L-Saccharopine	C11H20N2O6 C28H38N6O6 C11H20N2O6	MassBank Nist_msms2 Nist_msms	7.79 6.28 7.79
241	251.1499	C12H19N4O2	409	Trp-His-Arg	C23H31N9O4	Nist_msms	44.16
241	195.0876	C8H11N4O2	511	Cys-His-Lys	C15H26N6O4S	Nist_msm	22.64

				1,3-Dimethylurate	C7H8N4O3	s MassBank	7.68
241	157.0738	C7H11NO3	621	Co-elution and Co-fragmentation	-	-	-
241	272.0876	C10H16N3O6	592	5-Androsten-3.beta.,17.beta.-diol Androsterone	C19H30O2 C19H30O2	Nist_msms MassBank	18.48 6.06
241	363.1760	?	904	-	-	-	-
17	**293.1131**	C14H17N2O5	431	PyroGlu-Tyr Insulin-Like Growth [Tyr0] Factor II (33-40)	C14H16N2O5 C47H83N21O14	Nist_msms Nist_msms2	97.81 0.12
17	**182.0812**	C9H12NO3	585	Etilefrine L-Tyrosine	C10H15NO2 C9H11NO3	Nist_msms MassBank	76.41 6.67
17	280.1543	C15H22NO4	255	Tyr-Val	C14H20N2O4	Nist_msms	65.96
17	308.1856	C17H26NO4	234	DL-Octopamine Tyr-Met-Arg-Phe-NH2|1/3,F,Amidated|38/38	C8H11NO2 C29H42N8O5S	Nist_msms Nist_msms2	95.56 2.42
17	161.0921	C6H13N2O3	385	Bethanechol cation L-2-Aminoadipic acid	C7H17N2O2 C6H11NO4	Nist_msms MassBank	62.46 15.44
17	154.0974	C8H12NO2	417	5-Aminosalicylic acid 3-Sulfino-L-alanine	C7H7NO3 C3H7NO4S	Nist_msms MassBank	62.64 3.14
202	**205.0972**	C11H13N2O2	554	L-Tryptophan Trp	C11H12N2O2 C11H12N2O2	Nist_msms MassBank	98.99 98.99
202	**206.0811**	C11H12NO3	414	DL-Indole-3-lactic acid	C11H11NO3	Nist_msms	98.23
202	367.1500	C17H23N2O7	504	Trp(Dioxidation)-Glu	C16H19N3O7	Nist_msms	61.69
202	218.0811	C12H12NO3	279	N-Acetyl-5-hydroxytryptamine	C12H14N2O2	Nist_msms	64.65
202	334.1398	C16H20N3O5	541	Trp-Lys Trp-Glu	C17H24N4O3 C16H19N3O5	Nist_msms Nist_msm	97.37 0.31

						s	
202	188.0706	C11H10NO2	553	- (fragment)	-	-	-
202	291.0973	C14H15N2O5	561	(+)-Catechin	C15H14O6	Nist_msms	93.7
202	222.1124	C12H16NO3	288	2,6-Di-tert-butylbenzoquinone	C14H20O2	Nist_msms	71.83
202	277.1585*	C34H43N4OP	427	L-Saccharopine L-Saccharopine	C11H20N2O6 C11H20N2O6	MassBank Nist_msms	7.79 7.79
202	277.1474	?	436	Co-elution and Co-fragmentation	-	-	-
202	237.0869	C11H13N2O4	470	Carbetamide	C12H16N2O3	Nist_msms	73.34
202	208.0597	C10H12NO4	445	L-Kynurenine Kynurenine	C10H12N2O3 C10H12N2O3	Nist_msms MassBank	32.88 2.87
202	261.0934	?	508	Few low abundant fragments related to indole	-	-	-
202	190.1437 (**190.0861**)	C9H20NO3 (C11H12NO2)	435	1H-Indole-2-carboxylic acid, ethyl ester	C11H11NO2	Nist_msms	68.25
202	146.0599	C9H8NO	410	Fragment of 3-Indolepropionic acid	-	-	-

S2.5 Co-occurrences of Fragments and Losses in Matched Mass2Motifs from Different Samples

The correspondence of different Mass2Motifs, discovered through running MS2LDA independently on each beer sample, can be established through matching of the fragment or loss features that comprise the Mass2Motifs. Figure S-10 shows the same histidine-related Mass2Motifs discovered through explorations of the Beer1 and Beer3 results via MS2LDAVis. The 'Mass2Motif Feature Frequencies' histograms (Figure S-6A, S-6C) display how often particular fragments or losses appear in spectra including this Mass2Motif, indicating their consistency. For example, from Figure S-9A and S-9C we can see that the fragments 110.0718 ($[C_5H_8N_3]^+$) and 93.0450 ($[C_5H_5N_2]^+$) m/z are most consistently present in the histidine Mass2Motifs for Beer 1 and Beer 3. The 'Mass2Motif Global Frequencies' histograms (Figure S-9B, S-9D) show how specific these fragments and losses are to this Mass2Motif. The blue bars show the total abundance of each fragment (or loss) in the entire dataset whilst the red bars show the abundance that can be attributed to this Mass2Motif. We see from Figures S-6B and S-6D that globally, most of the observed fragments with m/z 110.0718 ($[C_5H_8N_3]^+$) are explained by these histidine-related Mass2Motifs, and whereas the fragment at m/z 95.0608 is consistently present in these Mass2Motifs, it is also abundantly present elsewhere.

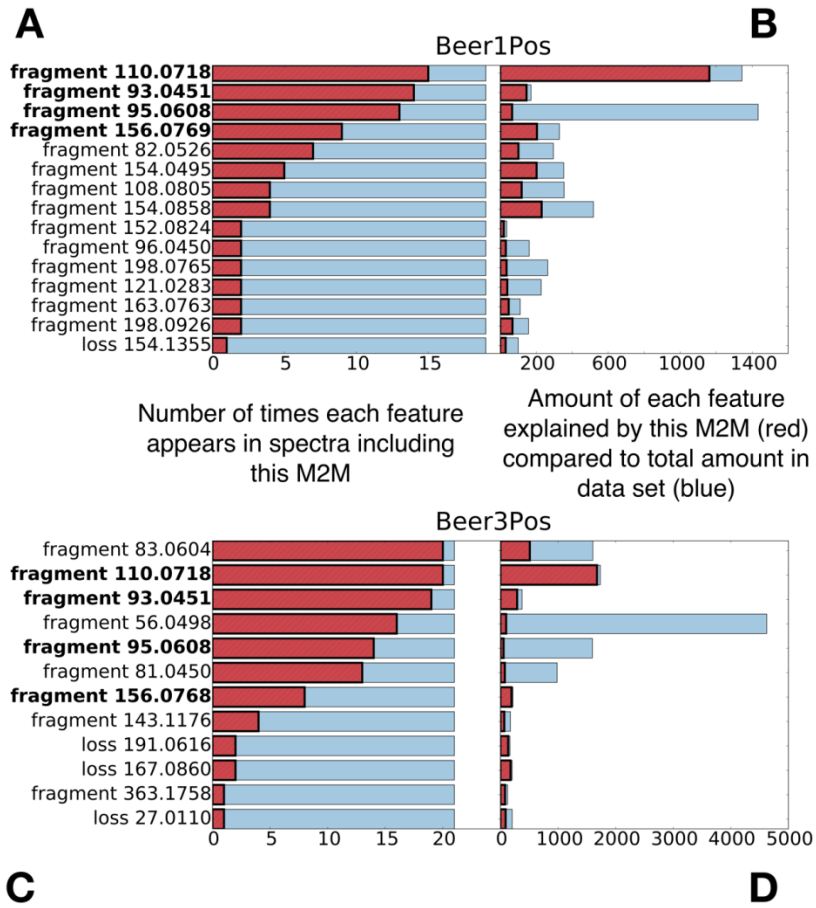


Figure S-10. Similar sets of fragment and loss features can be seen in the MS2LDAVis Feature Frequency histograms for the histidine-related Mass2Motifs in positive mode of Beer1 (top) and Beer3 (bottom). The left-hand panels of **A)** and **C)** show the number of times each feature appears in spectra associated with this Mass2Motif while the right-hand panels of **B)** and **D)** show the proportion (red) of the total abundance (blue) of this feature within the dataset explained by this Mass2Motif. Using **B)** as an example, we see that this Mass2Motif accounts for the vast majority of the total abundance observed for the fragment with mass 110.0718 in Beer1. Conversely, we also see in **B)** that although the fragment with mass 95.0608 appears often in the spectra associated with this Mass2Motif, it appears widely elsewhere too. Because the analyses of the four beers were done separately, fragment masses do not exactly match across samples.

S2.6 Similar yet Different Aromatic Substructures of Phenylethene, Ethylphenol, and Phenylethyleneamine

The following three aromatic substructures (illustrated in Figure S-11) were present and could be annotated to Mass2Motifs found in all positive ionization mode Beer files:

- Phenylethene
- Proposed aromatic substructure derived from cinnamic acid (cinnamate)
- [phenylalanine-CHOOH] or 1-(phenylethene)-amine.

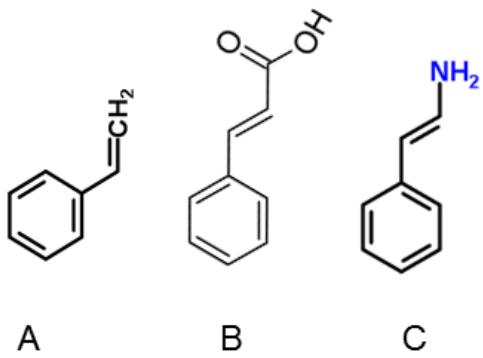


Figure S-11. Three aromatic substructures annotated to Mass2Motifs found in all four Beer fragmentation files, with **A)** phenylethene, **B)** proposed aromatic substructure derived from cinnamic acid (cinnamate), and **C)** [phenylalanine-CHOOH] or 1-(phenylethene)-amine.

Using the Beer2 positive ionization mode data as an example, the following list of Mass2Motifs is observed:

- **Mass2Motif 72** is a Phenylethene substructure motif. It has a degree of 38 and is characterized by the following fragment/loss features: fragment_105.0702 (C8H9), fragment_79.0541 (C6H7), **fragment_91.0541 (C7H7)**, fragment_53.0388, C8H9 (C4H5).

- **Mass2Motif 202** is a Cinnamic acid (cinnamate)-based substructure motif. It has a degree of 22 and is characterized by the following fragment/loss features: **fragment_91.0541 (C7H7)**, fragment_119.0488 (C8H7O), fragment_147.0437 (C9H7O2), fragment_65.0388 (C5H5).

- **Mass2Motif 117** is a [phenylalanine-CHOOH]-based substructure motif. It has a degree of 50 amd is characterized by the following fragment/loss featues: fragment_118.0654 (C8H8N), fragment_117.0571 (C8H7N), **fragment_91.0541 (C7H7)**, fragment_130.0645 (C9H8N), fragment_188.0706 (C11H10NO2).

In the above list, all Mass2Motifs share one fragment, highlighted in red, which is related to the aromatic core (mono-substituted benzene ring), i.e., fragment C_7H_7 $[M+H]^+$ (91.0541 m/z); however, in combination with other mass fragments, these three aromatic substructures are distinguishable by MS2LDA.

S2.7 Structurally Annotated Mass2Motifs Can Explain Matched Standards

The following list describes the Mass2Motifs, alongside their annotations, which can be associated to the fragmentation spectra of the Standard peaks shown in Figure 3 of the paper. The degree of a Mass2Motif indicates the number of MS2 fragmentation spectra in the beer3 positive ionization mode data having fragment or loss features that can be explained by the Mass2Motif (at the specified thresholding level).

- **Mass2Motif 115** is a [phenylalanine-CHOOH]-based substructure motif. It has a degree of 28 and is characterized by the following fragment/loss features: fragment_120.0808 (C8H10N), fragment_103.0546 (C8H7), fragment_91.0541 (C7H7).

- **Mass2Motif 156** is a [ribose (pentose, C5-sugar)-H2O]-related loss motif. It has a degree of 22 and is characterized by by the following fragment/loss features: loss_132.0421 (C5H8O4).

- **Mass2Motif 202** is a [tryptophan-NH3]-related substructure. It has a degree of 15 and is characterized by the following fragment/loss features: fragment_118.0654 (C8H7N), fragment_117.0571 (C7H7), fragment_91.0541 (C9H8N), fragment_130.0645 (C9H8N), fragment_188.0706 (C11H10NO2)

- **Mass2Motif 211** is an N-acetylputrescine substructure motif. It has a degree of 24 and is charactererized by the following fragment/loss features: loss_59.0370 (C2H5NO), fragment_114.0912 (C6H12NO), fragment_72.0447 (C3H6NO), fragment_60.0448 (C2H6NO).

- **Mass2Motif 214** is an amine loss motif. It has a degree of 57 and is characterized by the following fragment/loss features: loss_17.0247 (NH3).

- **Mass2Motif 220** is an adenine substructure motif. It has a degree of 32 and is characterized by the following fragment/loss features: fragment_136.0629 (C5H6N5), fragment_119.0351 (C5H3N4)

- **Mass2Motif 241** is a histidine substructure motif. It has a degree of 21 and is characterized by the following fragment/loss features: fragment_110.0718 (C5H8N3), fragment_156.0769 (C6H10N3O2), fragment_93.0450 (C5H5N2), fragment_95.0608 (C5H7N2).

- **Mass2Motif 262** is a combined loss of H2O and CO motif, indicative for free carboxylic acid group (COOH). It has a degree of 90 and is characterized by the following fragment/loss features: loss_46.0053 (CH2O2).

S2.8 GNPS and Massbank Results

To evaluate and validate the discovered Mass2Motifs using Beer fragmentation files, we performed MS2LDA analyses of the MassBank (10) and the Global Natural Products service (GNPS) (11) data sets as used by Dührkop et al. to train and test their CSI:FingerID tool (12). These datasets contain fragmentation spectra of thousands of reference compounds from different sources as chemical standards or isolated natural products. The fragmentation spectra were all acquired in positive ionization mode and generated at different instruments across the world. In (12), spectra from Orbitrap instruments were omitted (which allowed us to also test the extent to which Mass2Motifs are transferable across measurement platforms). A special feature extraction pipeline was developed to successfully bin mass fragments and losses from the diverse set of fragmentation spectra (see Section S1.1 for details). For LDA inference, Variational Bayes inference was applied to both data sets (see Section S1.2 for details). The resulting 1953 and 5670 spectra from MassBank and GNPS, respectively, were decomposed into 500 Mass2Motifs each.

Validation of beer-characterized Mass2Motifs in MassBank and GNPS data sets

To assess how well Mass2Motifs characterized in another dataset can be used for metabolite annotation in another dataset, the ~30 Mass2Motifs structurally characterized in beer were incorporated into the model (see Section S1.2), whilst the remaining Mass2Motifs were inferred by MS2LDA. To match the beer Mass2Motifs, we searched for the chemical formulas of the relevant fragments and neutral losses in the GNPS and Massbank features. A Mass2Motif was incorporated into the analysis if features corresponding to at least 50% of the Mass2Motifs probability could be found in GNPS or Massbank. This resulted in slightly different Mass2Motifs being matched in the two datasets (some beer features did not exist in the GNPS and Massbank data set) but a set of 22 Mass2Motifs were found in both. This demonstrates that the patterns of fragment and loss features that comprise Mass2Motifs can be transferred across spectra from different instruments.

As all the fragmented metabolite structures from the MassBank and GNPS datasets are known, we could validate the presence of beer-characterized Mass2Motif chemical substructures or chemical features in the molecular structures of spectra associations to these beer-characterized Mass2Motifs. Using 2D chemical structure images from ChemSpider (extracted via a search on InCHiKey using ChemSpiPy http://chemspipy.readthedocs.io/en/latest/) JvdH manually validated annotations on all molecules that included one or more of the beer Mass2Motifs by checking if the characterized substructure or structural feature were present in the molecular structures. In some cases, closely related substructures (not discriminable by mass spectrometry) were also considered as true, for example in case of isomeric substructures. In samples from one biological origin, substructures often relate to one isomer; however, in a set of thousands of standards, there is often more diversity. The resulting Tables for these analysis (GNPS_Mass2Motif_validations.csv and MassBank_Mass2Motif_validations.csv) can be found here: http://dx.doi.org/10.5525/gla.researchdata.313 .

From this manual validation, we computed two performance measures, the proportion of correct annotations at a probability threshold of 0.1 (i.e. if the spectra to Mass2Motif probability was >=0.1) with the results that 81.5% of annotations were correct in MassBank and 63.3% in GNPS. This shows the application of MS2LDA on different type of fragmentation spectra (other instruments), and the set of standards from MassBank and GNPS allowed us to determine false positive rates for the discovery of common substructures/structural features by MS2LDA. To investigate the performance across the different Mass2Motifs, we computed the Area Under the ROC curve for molecules connected to each Mass2Motif. The results are shown in Figure S-12. In a small number of Mass2Motifs, either all of the annotated molecules were correct, or all were incorrect making it impossible to define an AUC value. In these cases, we have instead plotted the accuracy at a threshold of 0.1. These cases are: Massbank: Mass2Motif 19, all incorrect but with probabilities below 0.1, Mass2Motif 20, all incorrect but all below 0.1 and GNPS: Mass2Motif 21, all incorrect but all with probabilities below 0.1.

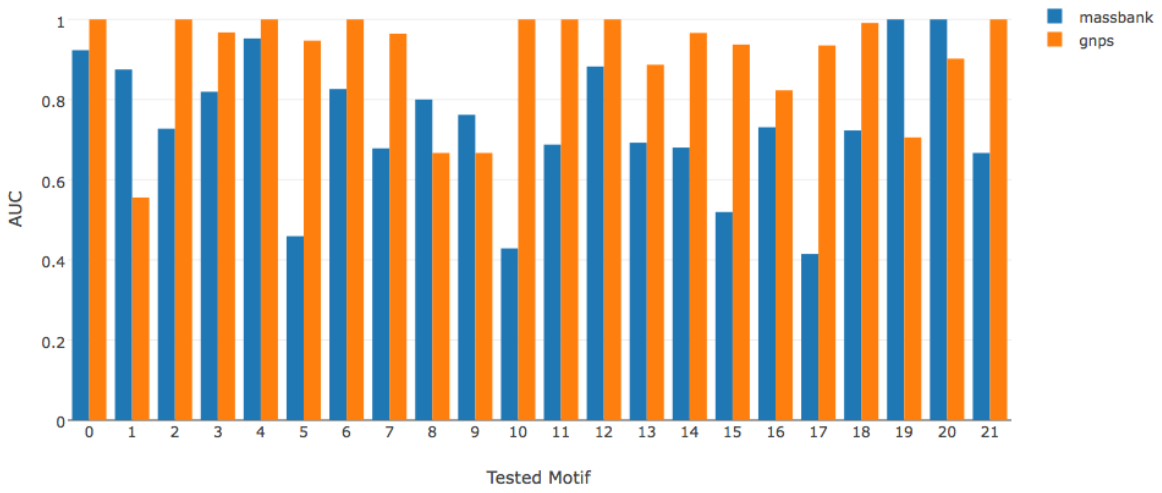


Figure S-12: GNPS and Massbank performance for the tested motifs. Motif numbers correspond to those in Table S-13. Motif names are shown in the table below.

Verified Mass2Motif	**Description**
0	Small nitrogen containing fragment ion (often proline or ornithine derived) most abundant fragment in beer data.
1	Fragments indicative for asparagine substructure (MzCloud), prevalent in Beer 3.
2	Oxygen-rich losses and fragments also occurring in hexose spectra - related to M2M 211 (hexose [glucose] conjugatation) - possibly hydrated-hexose loss?
3	Combined loss of H2O and CO, indicative for free carboxylic acid group (COOH) , a generic substructure in amino acids and organic acids.
4	Nitrogen containing substructure [C5H12N] (in beer related to Leucine).
5	Alkyl aromatic substructure - indicative for aromatic ring with 2-carbon alkyl chain attached i.e. phenylethene fragment from ethylbenzene as a result of the fragmentation process.
6	Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment).
7	[Pentose (C5-sugar)-H2O] related loss , indicative for conjugated pentose sugar - EF fits.
8	Fragment ions indicative for pyroglutamic acid (pyroglutamate) or glutamine (both in MzCloud) - structure can be formed from glutamic acid (glutamate) in the mass spectrometer as well.
9	Fragments indicative of a glycosylation , .e., indicative for a sugar conjugation (in beer often related to glucose).
10	Fragments indicative for histidine (C6H10N3O2) substructure (MzCloud)
11	Imidazole group linked to a carboxylgroup through one CH2 group i.e. like in imidazole acetic acid.

12	Fragment ions indicative for alkylamine substructure C5H10N (in beer often pipecolic acid [pipecolate]).
13	Fragments indicative for cinnamic/hydroxycinnamic acid substructure
14	Double water loss i.e. 2*H20 - Generic feature for metabolites containing several free OH groups attached to a aliphatic chain like sugars.
15	Water loss - indicative of a free hydroxyl group (in beer often seen in sugary structures).
16	Fragments indicative for [phenylalanine-CHOOH] based substructure.
17	CO loss - indicative for presence of ketone/aldehyde/lactone group (C=O).
18	Amine loss - Indicative for free NH2 group in fragmented molecule.
19	Fragment ions indicative for C6H12NO substructure (in beer related to N-acetylputrescine - MzCloud).
20	Fragments indicative for ferulic acid based substructure (MzCloud).
21	Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) - C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.

Table S-13: Characterisation of populated Mass2Motifs in GNPS and Massbank.

Assessment of number of validated Mass2Motifs per MassBank and GNPS fragmentation spectrum

MS2LDA can provide multiple annotations per molecule as multiple Mass2Motifs can be used to decompose an individual spectrum. Figure 4 in the manuscript demonstrates this for a single example. Here we investigate the extent to which the GNPS and Massbank molecules contain multiple validated beer Mass2Motif annotations. I.e., for each of the spectra with validated annotations, we count the number that have 1, 2, 3 or 4 validated annotations (i.e. to have 2 validated annotations, the molecule must include 2 of the Mass2Motifs structurally characterized in beer, both of which have been manually validated to be correct). The results can be seen in Figure S-14. In summary, of the 694 Massbank spectra that had one or more validated annotations, 173 had two, 36 3 and 3 4. For GNPS, of the 613 spectra with one or more, 34 had 2 and 4 had 3. In both cases, this demonstrates the large number of molecules for which MS2LDA can provide multiple annotations, thereby aiding in structural characterization. It is particularly noteworthy that all of this is from just the small number (~30) Mass2Motifs that we characterized from our beer analysis not including any MassBank or GNPS discovered Mass2Motifs.

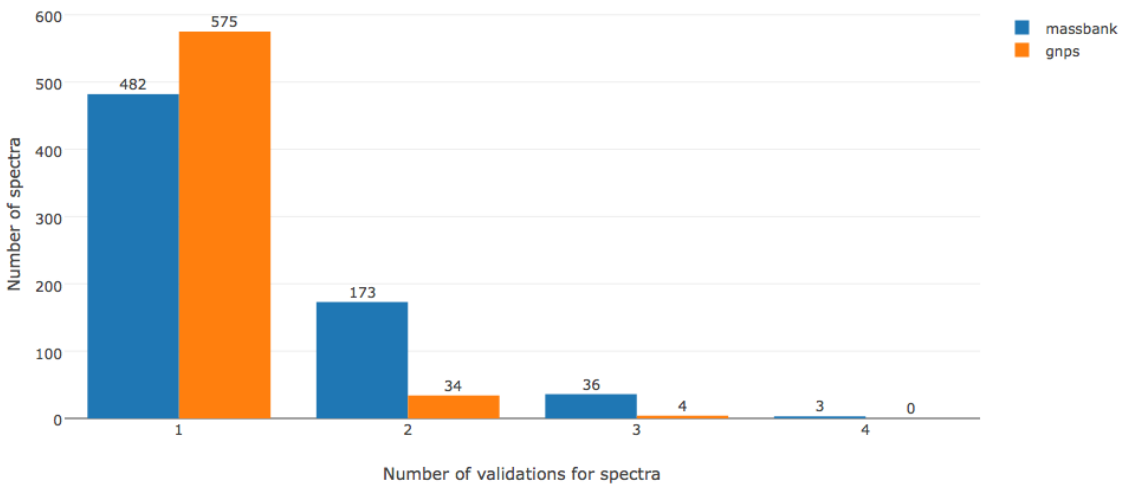


Figure S-14: the number of validations per spectra in the Massbank (blue) and GNPS (orange) data sets.

MS2LDA finds not previously characterized Mass2Motifs in MassBank and GNPS data sets

To assess whether MS2LDA could also discover new Mass2Motifs within the MassBank and GNPS data sets that were not previously characeterized in beer, we checked the resulting MS2LDA networks for non-beer-characterized mass2motifs, and we were able to structurally characterize 6 for each of the data sets to demonstrate the versatility of MS2LDA:

MassBank:

- motif_377: kaempferol/glycosylated kaempferol substructure (flavonoid – plant metabolite)
- motif_439: quercetin/glycosylated quercetin substructure (flavonoid – plant metabolite)
- motif_472: atenolol related (antihypertensive drug)
- motif_273: loss of [deoxyhexose-H2O]
- motif_377: loss of methyl group – indicative for presence of a methoxy [O-CH3] group
- motif_191: loss of C3H6 - indicative for the presence of an isopropyl group

GNPS:

- motif_214: benzene sulfonamide
- motif_176: 2-oxochromen-7-yl (mainly dimethylated)
- motif_436: 2-oxochromen-7-yl (mainly trimethylated)
- motif_121: sterone related
- motif_72: benzene chloride
- motif_287: C4H8 loss indicative for saturated C4-alkyl substructure (mainly tert-butylgroup and loss from 8,8-Trimethyl-2-oxo-9,10-dihydro-2H,8H-pyrano[2,3-f]chromen-5-yl substructure)

This indicates that MS2LDA can find a wide range of structurally diverse mass2motifs not related to the beer motifs, which are in fact complementary to those found in the beer data.

MS2LDA applied to urine data

MS2LDA was applied to fragmentation data from a human urine sample, representing a complex sample matrix (13). As with the GNPS and Massbank analyses, the structurally characterized Mass2Motifs from the beer analysis were incorporated through matching the relevant features. To validate the annotations provided by

these structurally characterized Mass2Motifs we detected the same 45 standard molecules that we were able to detect in the beer analysis via mass and RT matching (the urine sample was run in the same batch as the beer samples ensuring that only minimal RT drift had occurred). As the structural identify of these 45 molecules is known we manually validated the resulting annotations and found that at a threshold of 0.1, 74.3% of the annotations were validated. We also investigated the extent to which the same Mass2Motifs could be found in an analysis without them being fixed in the analysis a-priori. By matching features after processing and considering two Mass2Motif to match if shared features account for at least 0.5 of the probability in the Mass2Motif in both beer and urine, we found matches for 21 out of the 38 motifs structurally characterized in beer 3. These two analyses demonstrate the robustness of Mass2Motifs discovered through MS2LDA.

S2.9 Molecular Networking of Beer Fragmentation Files

To compare Molecular Networking with MS2LDA, the generated .mzXML files of the Beer fragmentation .RAW files were uploaded into the Global Natural Products Social Molecular Networking (GNPS) environment (http://gnps.ucsd.edu, a free account is required to log in) using FTP to transfer all the files and a text file containing information on the files as there are more than 6 different samples (files) that should be compared. Parameter optimization for molecular network generation for the high-resolution mass spectrometry data sets resulted in the following settings. The data was clustered with MS-Cluster with a precursor mass tolerance of 0.25 Da and a MS/MS fragment ion tolerance of 0.005 Da to create consensus spectra. Then, consensus spectra that contained less than 2 spectra were discarded. A network was created where edges were filtered to have a cosine score above 0.55 and 2 or more matched peaks. Further edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. The spectra in the network were then searched against GNPS' spectral libraries. The library's spectra were filtered in the same manner as the input data. All matches kept between network spectra, and the library's spectra were required to have a cosine score above 0.6 and at least 4 matched peaks. Analog search was enabled against the library with a maximum mass shift of 100.0 Da. Running times were under 10 minutes. The following list details all molecular networking parameters and their values used to generate the molecular networks used in the manuscript.

1. PAIRS_MIN_COSINE=0.55
2. ANALOG_SEARCH=1
3. tolerance.PM_tolerance=0.25
4. tolerance.Ion_tolerance=0.005
5. MIN_MATCHED_PEAKS=2
6. TOPK=10
7. CLUSTER_MIN_SIZE=2
8. MAXIMUM_COMPONENT_SIZE=120/100*
9. MIN_PEAK_INT=500.0
10. FILTER_STDDEV_PEAK_INT=2.0
11. RUN_MSCLUSTER=On
12. FILTER_PRECURSOR_WINDOW=0
13. FILTER_LIBRARY=1
14. WINDOW_FILTER=0
15. SCORE_THRESHOLD=0.6
16. MIN_MATCHED_PEAKS_SEARCH=4
17. MAX_SHIFT_MASS=100.0

For the MAXIMUM_COMPONENT_SIZE parameter, 120 was used for the positive ionization mode, and 100 for the negative ionization mode. These values were determined by starting at 80 and increase in steps of 20 till the largest network was smaller than the maximum component size.

Cytoscape, network visualization software, was used to further process and visualize the downloaded molecular network data. The recommended graphical layout style is FM3 which is available for Cytoscape versions 2.8.1 and below. Thus, the molecular network was uploaded into Cytoscape (version 2.8.1) following the

documentation available on the GNPS website. After applying the FM3 layout plugin, the molecular network was saved in .cys format (Cytoscape Session File) and reopened in Cytoscape version 3.2.0, where labelling and colouring of nodes and edges was conducted. Most importantly, the nodes were labelled with precursor masses, coloured using the rainbow pallet (two nodes having the same colour means that they are present in the same set of files, and accordingly, two nodes having similar colours means that they are present in a similar set of files, often differing in one or two files), and the size of the nodes was made proportional to the number of unique files from where the node spectra originated, i.e., the larger the node, the more unique files its spectra came from. The edges were labelled with the cosine similarity score of the two nodes they connect. The resulting molecular networks for both ionization modes were then inspected in the Cytoscape environment (see also (13)).

MS2LDA and Molecular Networking Comparison

Inspection of other clusters produced by Molecular Networking allowed us to identify clusters based on the core structures for histidine, tyrosine and tyramine (ethylphenol), as well as hydroxycinnamic acid, guanine and citric acid, in positive and negative ionization mode respectively. After a more detailed analysis of the Mass2Motifs related to ferulic acid, histidine, tyrosine, and tryptophan, we could annotate ferulic acid conjugates to polyamine structures like putrescine, histidine metabolites conjugated to hexose and organic acid moieties as well as a family of indole (tryptophan) related metabolites (see Supporting Information section 5.6 for more details). Two of those annotated beer metabolites were found to be dipeptides, whereas all others represent amino acids conjugated with other compound classes.

Based on the example shown in Figure 4 of the paper, it is likely that annotations of many molecules in these clusters could benefit from the flexibility of better decomposition of the spectra into multiple Mass2Motifs, rather than each parent ion having to be assigned to a single cluster alone. To illustrate with an example, we see in Figure S-15 a matrix of cosine similarities of some parent ions drawn from the ferulic acid based cluster and the tyramine based cluster constructed through molecular networking. We see clear, distinct groupings of these spectra into two clusters based on the parent ions' cosine similarities. Members of each cluster can therefore be explained by a single Mass2Motif (the ferulic acid cluster by M2M_19, and the tyramine cluster by M2M 58). However, one parent ion can also be explained by the two Mass2Motifs together. In cosine clustering, this parent ion would have to go into one cluster or the other based on its cosine similarity.

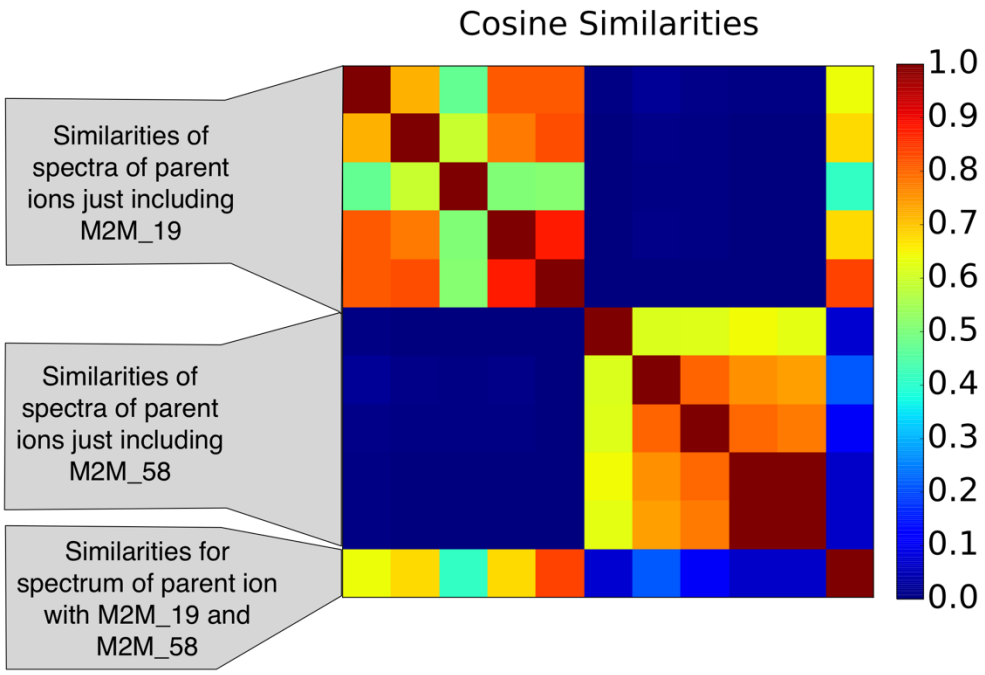


Figure S-15. Cosine clustering results of spectra drawn from the ferulic acid based cluster and the tyramine based cluster. The last row represents the spectrum containing both substructures, and is connected to one of the clustered based on cosine similarity scoring.

S2.10 Perplexity Comparison of MS2LDA and Multinomial Mixture Model

To validate the assumption of Mass2Motifs representing biological building blocks (i.e. fragmentation spectrum contains more than one Mass2Motifs), we compared the LDA model at the heart of the MS2LDA workflow to a multinomial mixture model that can be used for the clustering of fragmentation spectra (like Molecular Networking). The latter is equivalent to LDA with each spectrum being forced to consist of only one Mass2Motif. If MS2LDA is indeed finding structural features as conserved patterns of fragments and losses, it should explain the data with fewer Mass2Motifs than the mixture model. This is because the mixture model has to create separate Mass2Motifs for all observed combinations of structural features.

For model comparison, we plot perplexity (a measure of model fit; lower values indicate a better fit) for the two models as a function of K, the number of Mass2Motifs (for LDA) or clusters (for the mixture model). This is shown in Figure S-16. The lower perplexity values for $K > 100$ demonstrates that LDA provides a better model fit on the held-out data when compared to the mixture model, thus validating our assumption that allowing multiple conserved blocks to be present in small molecule fragmentation data is a better representation of the biochemical properties of the fragmented molecules. Details of the mixture models and on hyper-parameter optimizations and the cross-validation procedures of the two models are available in Section S1.2.

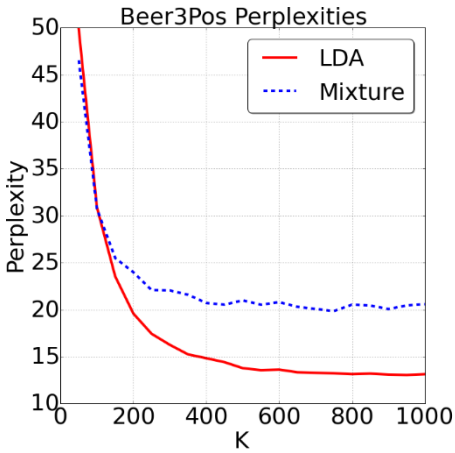


Figure S-16. Results of model comparisons of LDA and multinomial mixture model on the beer3 positive ionization mode dataset. The lower perplexity values for $K > 100$ demonstrates that LDA provides a better model fit on the held-out data when compared to the mixture model.

S2.11 Differential Analysis of Mass2Motifs

By linking the MS2LDA analysis with fold changes of MS1 peaks, we can assess the DE of Mass2Motifs, allowing us to identify biochemical changes across groups of samples based on which metabolites can be explained by a Mass2Motif. The advantage of this approach is for the purpose of differential analysis, there can more fragmentation spectra explainable by the MassMotifs in comparison to the number of spectra that can be annotated/identified through conventional means (see Discussion in the paper). This can be very useful, for example, in the case of a pathway-related Mass2Motif where we can assess the change in pathway activity across groups of samples without first having to identify and map molecules to the pathway.

For every Beer extract, LC-MS runs were processed using an in-house metabolomics pipeline (based on XCMS (1) and mzMatch(14)). Peak tables were exported to .csv files, and the linking of MS1 peaks in the MS2LDA analysis to the MS1 peaks in the exported peak tables was performed through a greedy matching scheme. For each MS1 peak in MS2LDA, we find its corresponding MS1 peak in the exported peak table within a specified

mass and RT tolerance values (3 ppm, 30 seconds). If there are multiple possible matches, the one with the nearest m/z difference is selected. Following this, for each Mass2Motif, we construct a matrix where each row is a linked MS1 peak that can be explained by that Mass2Motif and the columns are intensity values from the different case/control groups. This matrix is used as input to our implementation of PLAGE (15), the output of which are the PLAGE scores of differentially expressed Mass2Motifs.

Figure S-17 shows four examples of Mass2Motifs with high PLAGE scores, which we have annotated as related to guanine, tryptophan, tyrosine and pentose loss substructures (details on their MS1 peak annotations are in Table S-18). Comparing against spectral similarity clustering, the molecules explainable by the pentose Mass2Motif (Figure S-17D) are distributed over 10 spectral clusters. Similarly, the 9 tryptophan (indole) related metabolites (many of which are considerably more abundant in Beer 2 than Beer 3) that can be explained by the tryptophan Mass2Motif (Figure S-17B) were distributed over 7 spectral clusters.

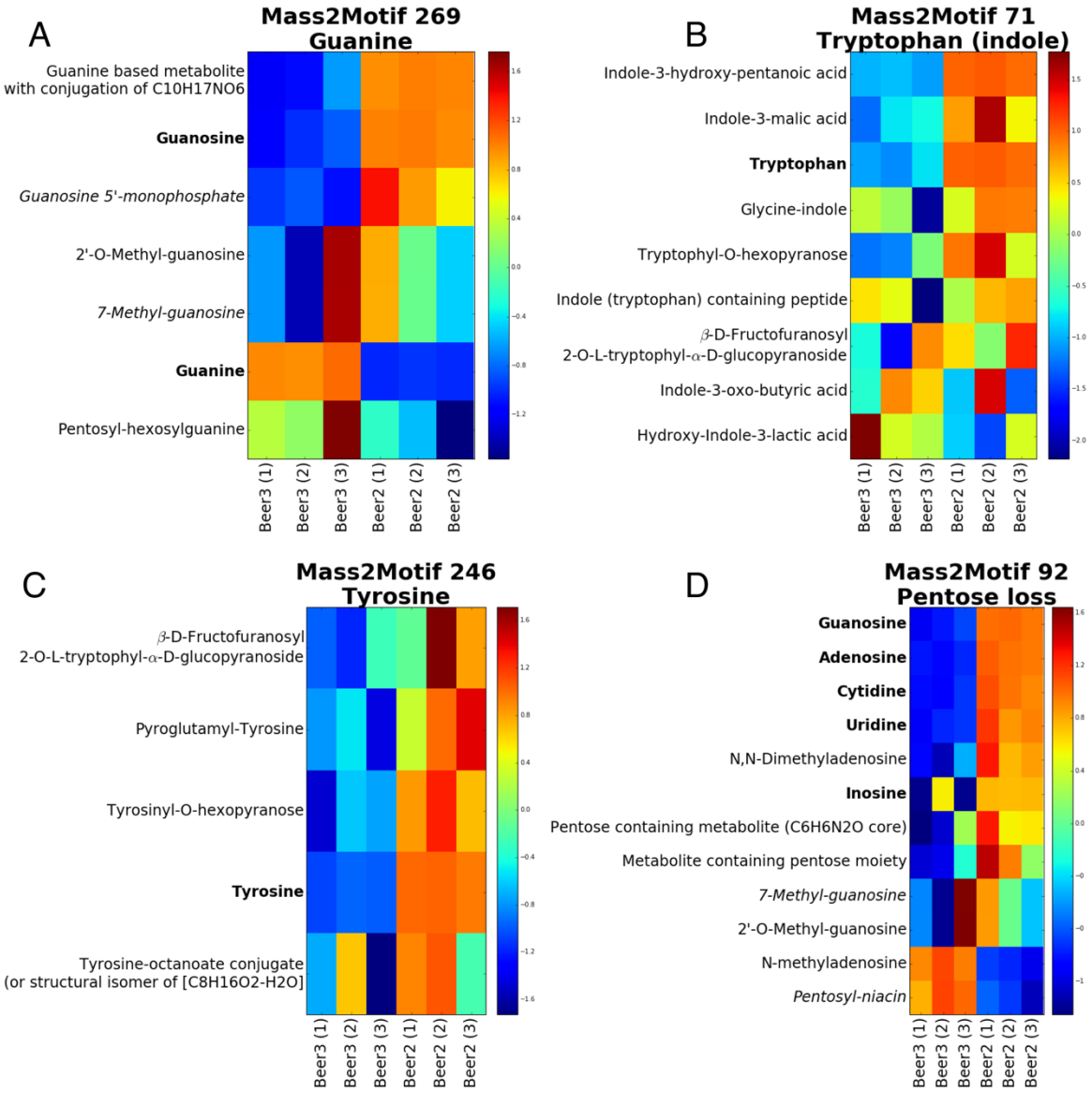


Figure S-17: Log fold change heat-maps for the A) guanine, B) tryptophan, C) tyrosine and D) pentose loss Mass2Motifs. Each row is an annotated MS1 peak and columns represent samples. For this validation, Metabolite identification was performed manually based on the Metabolite Standard Initiative Metabolite Identification scheme. Bold labels indicate identification at the highest level of confidence (1), while italic labels indicate identification at the next level of confidence (2). The remainder are level (3) or (4).

Table S-18. Annotation details on all MS1 peaks that can be explained by the four differentially-expressed Mass2Motifs in Figure S-17. All metabolites were annotated and validated from the Beer2 positive mode ionization data.

Mass $[M+H]^+$	**EF $[M+H]^+$ (most likely)**	**RT (s)**	**Class**	**Annotation**	**MSI MI level**
364.0651	C10H15N5O8P	625	Guanine	Guanosine 5'-monophosphate	2 (Nist; MassBank)
~~567.1912~~	~~Artefact~~	~~582~~	~~Guanine~~	~~Ion product of 284.0988~~	-
284.0988	C10H15N5O5	583	Guanine	Guanosine	1 (Nist; MassBank)
399.1623	C15H23N6O7	579	Guanine	Guanine based metabolite with conjugation of C10H17NO6	3
298.1146	C11H16N5O5	485	Guanine	2’-O-Methyl-guanosine	3 (mzCloud)
298.1146	C11H16N5O5	497	Guanine	7-Methyl-guanosine	2 (Nist; mzCloud)
152.0569	C5H6N5O	583	Guanine	Guanine	1 (Nist + standard)
446.1514	C16H24N5O10	675	Guanine	Pentosyl-hexosylguanine	3
205.0972	C11H13N2O2	554	Tryptophan (indole)	Tryptophan	1
~~205.1183~~	~~C8H11N2O4~~	~~597~~	~~Tryptophan (indole)~~	~~Co-fragmentation with Tryptophan~~	-
236.0916	C12H14NO4	399	Tryptophan (indole)	Hydroxy-Indole-3-lactic acid	3
425.1916	C21H25N6O4	298	Tryptophan (indole)	Indole (tryptophan) containing peptide? Co-fragmentation with isobars	4
217.0971	C12H13N2O2	522	Tryptophan (indole)	Glycine-indole? – Indole containing metabolite	4
218.0811	C12H12NO3	364	Tryptophan (indole)	Indole-3-oxo-butyric acid	3
367.1500	C17H23N2O7	504	Tryptophan (indole)	Tryptophyl-O-hexopyranose	3
236.1281	C13H18NO3	270	Tryptophan (indole)	Indole-3-hydroxy- pentanoic acid	3
252.0864	C10H12N4O4	414	Tryptophan (indole)	Indole-3-malic acid	3

529.2027	C23H33N2O12	546	Tryptophan (indole)	β-D-Fructofuranosyl 2-O-L-tryptophyl-α-D-glucopyranoside	3
~~262.1396~~	~~C10H20N3O5~~	~~487~~	~~Tryptophan (indole)~~	~~co-fragmentation with isobar containing indole~~	-
~~409.1869~~	~~Artefact~~	~~553~~	~~Tryptophan (indole)~~	~~Ion product of 205.1183~~	-
293.1131	C14H17N2O5	431	Tyrosine	Pyroglutamyl-Tyrosine	3
182.0812	C9H12NO3	585	Tyrosine	Tyrosine	1
308.1856	C17H26NO4	234	Tyrosine	Tyrosine-octanoate conjugate (or structural isomer of [C8H16O2-H2O]	3
~~182.0812~~	~~Artefact~~	~~610~~	~~Tyrosine~~	~~Shoulder peak of 182.0812~~	-
~~194.0811~~	~~C10H12NO3~~	~~362~~	~~Tyrosine~~	~~Noisy peak – not Tyrosine related (two fragments overlap) or fragment metabolite containing Tyrosine substructure~~	-
~~239.1123~~	-	~~509~~	~~Tyrosine~~	~~Co-fragmentation with isobars~~	-
506.1873	C21H32NO13	571	Tyrosine	β-D-Fructofuranosyl 2-O-L-tyrosinyl-α-D-glucopyranoside	3
344.1339	C15H22NO8	536	Tyrosine	Tyrosinyl-O-hexopyranose	3
~~378.1160~~	~~C15H22O11~~	~~592~~	~~Tyrosine~~	~~Not tyrosine related some fragments overlap~~	-
~~279.1547~~	~~C11H23N2O6~~	~~412~~	~~Tyrosine~~	~~Not tyrosine related some fragments overlap~~	-
268.1039	C10H14N5O4	469	Pentose loss	Adenosine	1
284.0988	C10H15N5O5	583	Pentose loss	Guanosine	1 (Nist; MassBank)
269.0879	C10H13N4O5	523	Pentose loss	Inosine	1
298.1146	C11H16N5O5	485	Pentose loss	2'-O-Methyl-guanosine (146.0723 loss is also part of Mass2Motif)	3 (mzCloud)
298.1146	C11H16N5O5	497	Pentose loss	7-Methyl-guanosine	2 (Nist; mzCloud)
446.1514	C16H24N5O10	675	Pentose loss	Pentosyl-hexosylguanine	3
282.1190	C11H16N5O4	675	Pentose	N-methyladenosine	3

			loss		
390.1520	C13H29NO10P	622	Pentose loss	Metabolite containing pentose moiety	4
244.0926	C9H14N3O5	558	Pentose loss	Cytidine	1
245.0767	C9H13N2O6	499	Pentose loss	Uridine	1
255.0973	C11H15N2O5	1070	Pentose loss	Pentose containing metabolite (C6H6N2O core)	3
256.0814	C11H14NO6	583	Pentose loss	Pentosyl-niacin	2 (mzCloud for niacin fragments)
296.1353	C12H18N5O4	405	Pentose loss	N,N-Dimethyladenosine	2

S2.12 MS2LDA Uses High-Resolution Mass Spectrometry Information in the MS2 Domain

High-resolution mass spectrometry results in accurate mass measurements, also of detected mass fragments in the smaller m/z range of 50 – 70 Da. While it is generally true that fragments below 70 Da are found in more different annotated motifs than those above 70 Da, we could observe 19 different fragments with a nominal mass of 70 or lower. In 6 cases, two of those fragments have the same nominal mass, and in 1 case even three fragments share the same nominal mass: 60.0448 (C2H6NO, [M+H]+), 60.0559 (CH6N3, [M+H]+), and 60.0810 (C3H10N, [M+H]+). This shows the importance of using accurate mass fragmentation data as input to enable distinction between those fragment sets, and other isobaric fragments of higher m/z. Some of these fragments are unique for a substructure, for example, for CH6N3 the guanidine group is the only likely formation of the atoms, especially taking biological extracts as samples into account. Others are more generic, i.e., C4H5 and C4H7, but are part of Mass2Motifs pointing to different structural features in combinations with mass fragments of higher m/z.

S2.13 Spectral Matching of Mass2Motifs Using Their Reconstructed Mass Spectra

Table S-19 shows the results from reconstructing fragmentation spectra from various Mass2Motifs discovered through MS2LDA (see for examples Figure S-20) and using them to perform spectral matching to the NIST MSMS (Nist_msms) and MassBank spectral databases. Reconstruction of the spectra was performed by taking into account all the fragment features above the user-defined threshold t_φ on the Mass2Motif-to-features distributions [the φ parameters]). Here, t_φ is set to 0.01, which is the same value used for visualisation in MS2LDAvis. The counts of fragment features from the data that can be explained the Mass2Motif are then converted into relative intensities. This shows the potential to automatically structurally characterize Mass2Motifs.

M2M	Database Annotations	EFs top hits of each database	Database	Score
13	**L-Glutamine**	C5H10N2O3	Nist_msms	94.38
17	**L-Tyrosine**	C9H11NO3	Nist_msms	78.95
19	**trans-Ferulic acid**	C10H10O4	Nist_msms	76.94
40	Gln-Gly-Lys	C13H25N5O5	Nist_msms	11.07
42	**L-Asparagine** **Asn**	C4H8N2O3 C4H8N2O3	Nist_msms MassBank	97.15 97.15
45	L-Lysine	C6H14N2O2	Nist_msms	71.93
55	4-Hydroxy**cinnamic acid** (L-phenylalanine methyl ester) amide 4-Coumaric acid (=4-hydroxycinnamic acid)	C19H19NO4 C9H8O3	Nist_msms MassBank	50.69 11.33
58	**Phenol, 4-(2-aminoethyl)** **(=Tyramine)**	C8H11NO	Nist_msms	75.33
67	**cis-Aconitic acid**	C6H6O6	Nist_msms	97.4
69	**D-(+)-Arabitol**	C5H12O5	Nist_msms	40.51
79	**Betaine** **Betaine**	C5H11NO2 C5H11NO2	Nist_msms MassBank	98.64 98.64
82	Guanidine, (4-aminobutyl)-	C5H14N4	Nist_msms	71.18
91	5-Aminosalicylic acid	C7H7NO3	Nist_msms	83.59
98	1-Aminocyclohexane-carboxylic acid L-2-Aminoadipic acid	C7H13NO2 C6H11NO4	Nist_msms MassBank	88.76 1.88
115	**2-Amino-1-phenylethanol** **(Phenylethanolamine)**	C8H11NO	Nist_msms	91.03

129	**Lactulose**	C12H22O11	Nist_msms	32.74
130	Uridine L-Asparagine	C9H12N2O6 C4H8N2O3	MassBank Nist_msms	58.13 17.1
131	**D-(+)-Cellobiose**	C12H22O11	Nist_msms	54.02
158	Gly-Leu	C8H16N2O3	Nist_msms	54.26
162	Acyclovir (acycloguanosine)	C8H11N5O3	Nist_msms	89.55
166	**5-Methylcytosine** **5-Methylcytosine**	C5H7N3O C5H7N3O	Nist_msms MassBank	52.69 52.69
174	L-Glutamic acid N-Acetylglutamate	C5H10N2O3 C7H11NO5	Nist_msms MassBank	15.89 10.27
184	Trimethylamine N-oxide	C3H9NO	Nist_msms	88.62
202	**L-Tryptophan**	C11H12N2O2	Nist_msms	72.71
211	**N-acetylputrescine** Guanidine, (4-aminobutyl)	C6H14N2O C5H14N4	MassBank Nist_msms	79.53 18.75
220	**.beta.-Nicotinamide adenine dinucleotide, reduced** **Adenosine**	C21H29N7O14P2 C10H13N5O4	Nist_msms MassBank	16.98 8.55
222	**L-Serine**	C3H7NO3	Nist_msms	95.01
226	15-Deoxy-.DELTA.12,14-prostaglandin D2	C20H30O4	Nist_msms	17.29
230	L-NG-Nitroarginine methyl ester	C7H15N5O4	Nist_msms	24.84
241	N-.alpha.-(tert-Butoxycarbonyl)-**L-Histidine** **L-Histidine**	C11H17N3O4 C6H9N3O2	Nist_msms MassBank	73.25 15.32
276	2,6-Xylidine	C8H11N	Nist_msms	88.45
284	**1,2,3-Benzenetriol**	C6H6O3	Nist_msms	91.25

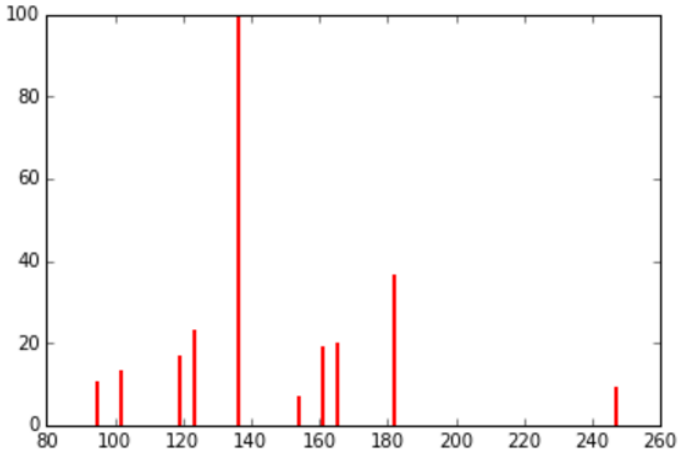
Table S-19. Reconstructed mass spectra from conserved patterns found in fragment-based Mass2Motifs searched in Nist and MassBank databases – top annotations for each database (if any) are indicated with their scores and highlighted in bold if they structurally matched manual annotations.

Figure S-20 shows examples of reconstructed Mass2Motifs of tyrosine (M2M_17), ferulic acid (M2M_19), 5-Methylcysteine (M2M_166) and histidine (M2M_241) related motifs. These reconstructed spectra were then used to search in the Nist_msms and MassBank libraries.

A) Mass2Motif 17 – Tyrosine related

Reconstructed MS2 peak list and mass spectrum:

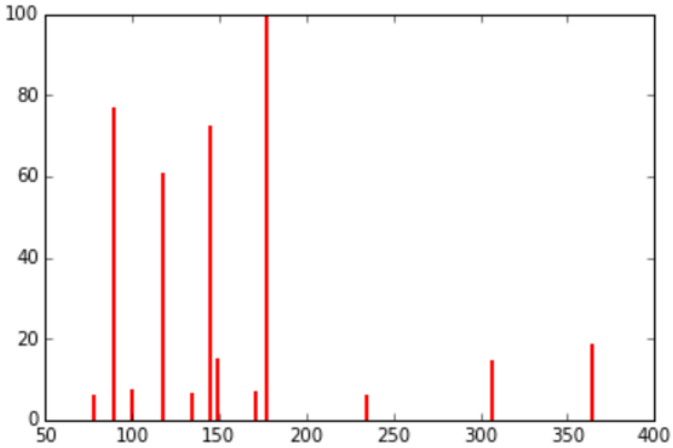
m/z [M+H$^+$]	Relative intensity
136.07599	100.0
182.08217	36.1
123.04467	22.7
165.05388	19.4
160.90206	18.6
119.04874	16.3
102.0547	12.7
95.04936	10.0
247.1084	8.65
161.0686	7.50
119.04991	6.93
165.05578	6.74
154.08575	6.54



B) Mass2Motif 19 – Ferulic acid related

Reconstructed MS2 peak list and mass spectrum:

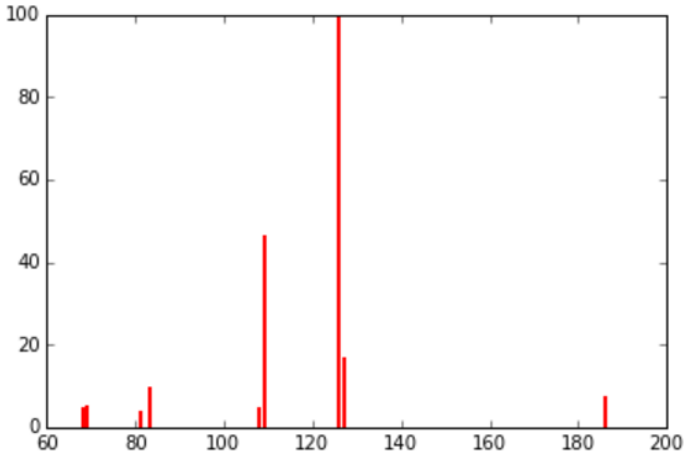
m/z [M+H$^+$]	Relative intensity
177.05475	100.0
89.03864	76.7
145.02839	72.1
117.03316	60.6
364.22203	18.4
149.05998	14.7
307.17496	14.4
100.07536	6.90
171.1487	6.52
134.03657	6.39
78.04655	5.89
234.11111	5.77



C) Mass2Motif 166 – 5-Methylcysteine related

Reconstructed MS2 peak list and mass spectrum:

m/z [M+H$^+$]	Relative intensity
126.0665	100.0
109.03967	46.1
127.03204	16.4
83.06041	9.52
186.10718	6.90
69.05759	4.77
68.04977	4.61
108.05597	4.28
81.04501	3.62



D) Mass2Motif 241 – Histidine related

Reconstructed MS2 peak list and mass spectrum:

m/z [M+H$^+$]	Relative intensity
110.07176	100.0
83.06041	29.1
93.04509	18.1
156.07684	12.7
56.04977	5.88
363.17581	5.88
143.11757	4.47
81.04501	4.41
95.06076	3.23

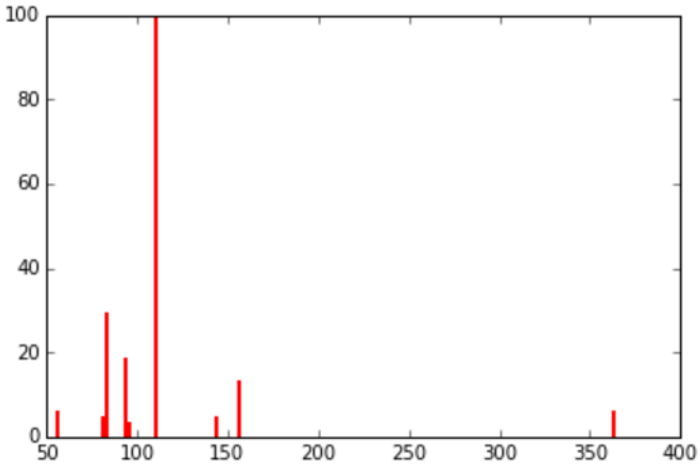


Figure S-20 Reconstructed mass spectra from Mass2Motifs found in beer data that could be used for spectral matching.

SECTION S3. BEER SAMPLES INFORMATION

Beer samples from three commercial beers and one home-brewed beer were used as representative complex mixtures of diverse biochemical:

- Beer1 is from a home-brewed bottle of German Wheat Beer (the Beer sheet can be found in S3.1 – S3.6 below).
- Beer2 is from a bottle of 'Jaw Glyde Ale' (a Golden/Blond Ale; http://www.jawbrew.co.uk).
- Beer3 is from a bottle of 'Seven Giraffes Extraordinary Ale' (an IPA style beer; http://www.williamsbrosbrew.com/beerboard/bottles/seven-giraffes).
- Beer4 is from a bottle of 'Black Sheep Ale' (a Golden Bitter Ale; https://www.blacksheepbrewery.com/beers/15/black-sheep-ale).

S3.1 General information

Type	German Wheat Beer - Weizen/Weissbier (15 A)
Type	All Grain
Batch Size	19.00 l
Boil Size	27.97 l
Boil Time	60 min
End of Boil Vol	23.70 l
Final Bottling Vol	16.16 l
Fermentation	Ale, Single Stage
Date	02 Jan 2015
Brewer	Paul Simon
Equipment	Paul's Kit
Efficiency	50.00 %
Est Mash Efficiency	60.0 %
Taste Rating	30.0

S3.2 Ingredients

#	Name	Type	Amt	%/IBU
1	White Wheat Malt (4.7 EBC)	Grain	4075.88 g	53.7%
2	Pale Malt (2 Row) UK (5.9 EBC)	Grain	3000.0 g	39.5 %
3	Munich Malt (17.7 EBC)	Grain	335.0 g	4.4%
4	Melanoiden Malt (39.4 EBC)	Grain	113.0 g	1.5%
5	Caramel/Crystal Malt - 40L (78.8 EBC)	Grain	40.0 g	0.5%
6	Chocolate Malt (689.5 EBC)	Grain	28.00 g	0.4%
7	Hallertauer Hersbrucker [4.00 %] - Boil 60.0 min	Hop	30.66 g	13.4 IBUs
8	Hallertauer Hersbrucker [4.00 %] - Boil 15.0 min	Hop	17.01 g	3.7 IBUs

S3.3 Gravity, Alcohol Content and Color

Est Original Gravity	1.063 SG
Est Final Gravity	1.016 SG
Estimated Alcohol by Vol	6.2 %
Bitterness	17.1 IBUs
Est Color	17.1 EBC
Measured Original Gravity	1.070 SG
Measured Final Gravity	1.020 SG
Actual Alcohol by Vol	6.6 %

Calories	675.7 kcal/l

S3.4 Mash Profile

Mash Name	Single Infusion, Medium Body
Sparge Water	4.69 l
Sparge Temperature	75.6 C
Adjust Temp for Equipment	TRUE
Total Grain Weight	7591.88 g
Grain Temperature	20.0 C
Tun Temperature	20.0 C
Mash PH	5.20

S3.5 Mash Steps

Name	Description	Step Temperature	Step Time
Mash In	Add 20.88 l of water at 75.5 C	66.7 C	60 min
Mash Out	Add 11.09 l of water at 95.8 C	75.6 C	10 min
Sparge	Fly sparge with 4.69 l of water at 75.6 C		

Mash Notes: Simple single infusion mash for use with most modern well modified grains (about 95% of the time).

S3.6 Carbonation and Storage

Carbonation Type	Bottle
Pressure/Weight	110.11 g
Keg/Bottling Temperature	21.1 C
Fermentation	Ale, Single Stage
Volumes of CO2	2.7
Carbonation Used	Bottle with 110.11 g Table Sugar
Age for	30.00 days
Storage Temperature	18.3 C

SECTION S4. DATA ACQUISITION WORKFLOW

Blank runs, quality control samples, and 3 standard mixes containing 150 reference compounds were run to assess the quality of the mass spectrometer and aid in metabolite annotation and identification (16). The pooled sample was run prior to and across the batch to monitor the stability and quality of the LC-MS run, whereas the samples were run in a randomized order. Immediately after acquisition, all .raw files were converted into MzXML format, thereby centroiding the mass spectra and separating positive and negative ionization mode spectra into two different mzXML files using the command line version of MSconvert (ProteoWizard). Fragmentation files were also converted into .mzML formats using the GUI version of MSconvert.

Accurate masses of standards were obtained well within 3 ppm accuracy and intensities of the quality control samples (a beer extract and a serum extract) were as expected. Six runs were collected for each beer sample, as well as the pooled beer sample, so that three combined full scan mode files were recorded, one combined fragmentation mode file, and two separate fragmentation mode files, one for (+) and one for (-) mode.

SECTION S5. MS AND MS/MS SETTINGS

S5.1 Positive Negative Ionization Combined Fragmentation Mode

A duty cycle consisted of a full scan in positive ionization mode, followed by a TopN data dependent MS/MS (MS2) fragmentation event taking the 10 most abundant ion species not on the dynamic exclusion list, followed by the same two scan events in negative ionization mode. Data acquisition was carried out in positive (+) and negative (-) switching ionization mode, using m/z 74.0964 (+) (ACN cluster), 88.07569 (+) (contaminant), and m/z 112.98563 (-) (Formic Acid cluster) as locking masses. The set up was calibrated [Thermo calmix, with additional masses at lower m/z; 74.0964 m/z (+) and 89.0244 (-)] in both ionization modes before analysis and a tune file targeted towards the lower m/z range was used.

In both ionization modes full scan (MS1) data was acquired in profile mode at 35,000 resolution using 1 microscan, an AGC target of 1E6 cts, a maximum injection time of 120 milliseconds, with spray voltages +3.8 kV (+) and −3.0 kV (-), probe heater temperature 150 °C, capillary temperature 320 °C, sheath gas flow rate 40, auxiliary gas flow rate 15 a.u., sweep gas flow rate 1 a.u, and a full scan mass window of 70–1050 m/z.

MS/MS (data dependent-MS2) data was acquired in profile mode at 35,000 resolution using 1 microscan, an AGC target of 1E5 cts, a maximum injection time of 120 milliseconds, a loop count of 10, a MSX count of 1, a TopN of 10, an isolation window of 1.0 Da, an isolation offset of 0.0 Da, a stepped normalized collision energy (NCE) higher collision dissociation (HCD) mode combining 25.2, 60.0, and 94.8 NCEs into one fragmentation scan, an undefill ratio of 20%, an intensity threshold of 1.7E5 cts, and the dynamic exclusion was set to 15 seconds. These settings result in a maximum duty cycle time (with two full scans and 20 MS2 scans) of 2.64 seconds, whilst in practice cycle times are shorter as not all 10 MS2 scans are always recorded or the ACG target was reached prior to the maximum filling time. Further settings were: no apex trigger, no charge exclusion, peptide match was off, exclude isotopes was on, and if idle, the machine did not pick up other ions.

S5.2 Positive or Negative Ionization Separate Fragmentation modes

As for the combined files, with the following modifications: full scan (MS1) resolution was set to 70,000, MS/MS (MS2) resolution was set to 17,500, MS/MS maximum injection time was set to 80 milliseconds and the undefill ratio set to 10%, with a resulting intensity threshold of 1.3E5 cts. The duty cycle consisted of one full scan (MS1) event and one Top10 MS/MS (MS2) fragmentation event. These settings result in a maximum duty cycle time (with one full scan and 10 MS2 scans) of 920 milliseconds, whilst in practice cycle times are shorter as not all 10 MS2 scans are always recorded or the ACG target was reached prior to the maximum filling time.

REFERENCES

1. Smith CA, Want EJ, O'Maille G, Abagyan R, & Siuzdak G (2006) XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* 78(3):779-787.
2. Stravs MA, Schymanski EL, Singer HP, & Hollender J (2013) Automatic recalibration and processing of tandem mass spectra using formula annotation. *Journal of Mass Spectrometry* 48(1):89-99.
3. Griffiths TL & Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228-5235.
4. Wallach HM, Murray I, Salakhutdinov R, & Mimno D (2009) Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning* 4:1105–1112.
5. Blei DM, Ng AY, & Jordan MI (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993-1022.
6. Böcker S, Letzel MC, Lipták Z, & Pervukhin A (2009) SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* 25(2):218-224.
7. Böcker S & Lipták Z (2007) A Fast and Simple Algorithm for the Money Changing Problem. *Algorithmica* 48(4):413-432.
8. Kind T & Fiehn O (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *Bmc Bioinf.* 8:art. no. 105.
9. Sievert C & Shirley KE (2014) LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*:63-70.
10. Horai H, *et al.* (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45(7):703-714.
11. Yang JY, *et al.* (2013) Molecular Networking as a Dereplication Strategy. *Journal of Natural Products* 76(9):1686-1699.
12. Dührkop K, Shen H, Meusel M, Rousu J, & Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences* 112(41):12580-12585.
13. van der Hooft JJJ, Padmanabhan S, Burgess KEV, & Barrett MP (2016) Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics* 12(7):1-15.
14. Scheltema RA, Jankevics A, Jansen RC, Swertz MA, & Breitling R (2011) PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Analytical Chemistry* 83(7):2786-2793.
15. Tomfohr J, Lu J, & Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *Bmc Bioinformatics* 6(1):1-11.
16. Creek DJ, *et al.* (2011) Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Analytical Chemistry* 83(22):8703-8710.